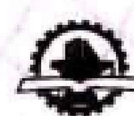


"十二五"规划教材

概率统计与SPSS应用

(第2版)

于义良 罗蕴玲 安建业 编著

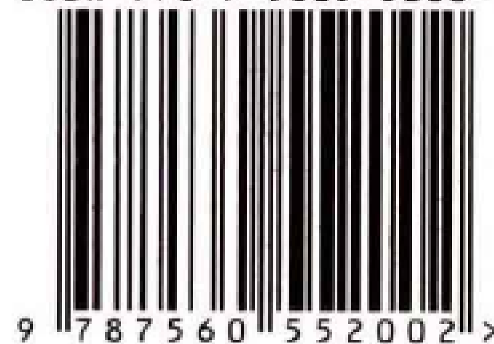


西安交通大学出版社
XI'AN JIAOTONG UNIVERSITY PRESS



"十二五"规划教材




ISBN 978-7-5605-5200-2

责任编辑：任振国　装帧设计：伍　胜

定价：27.00元

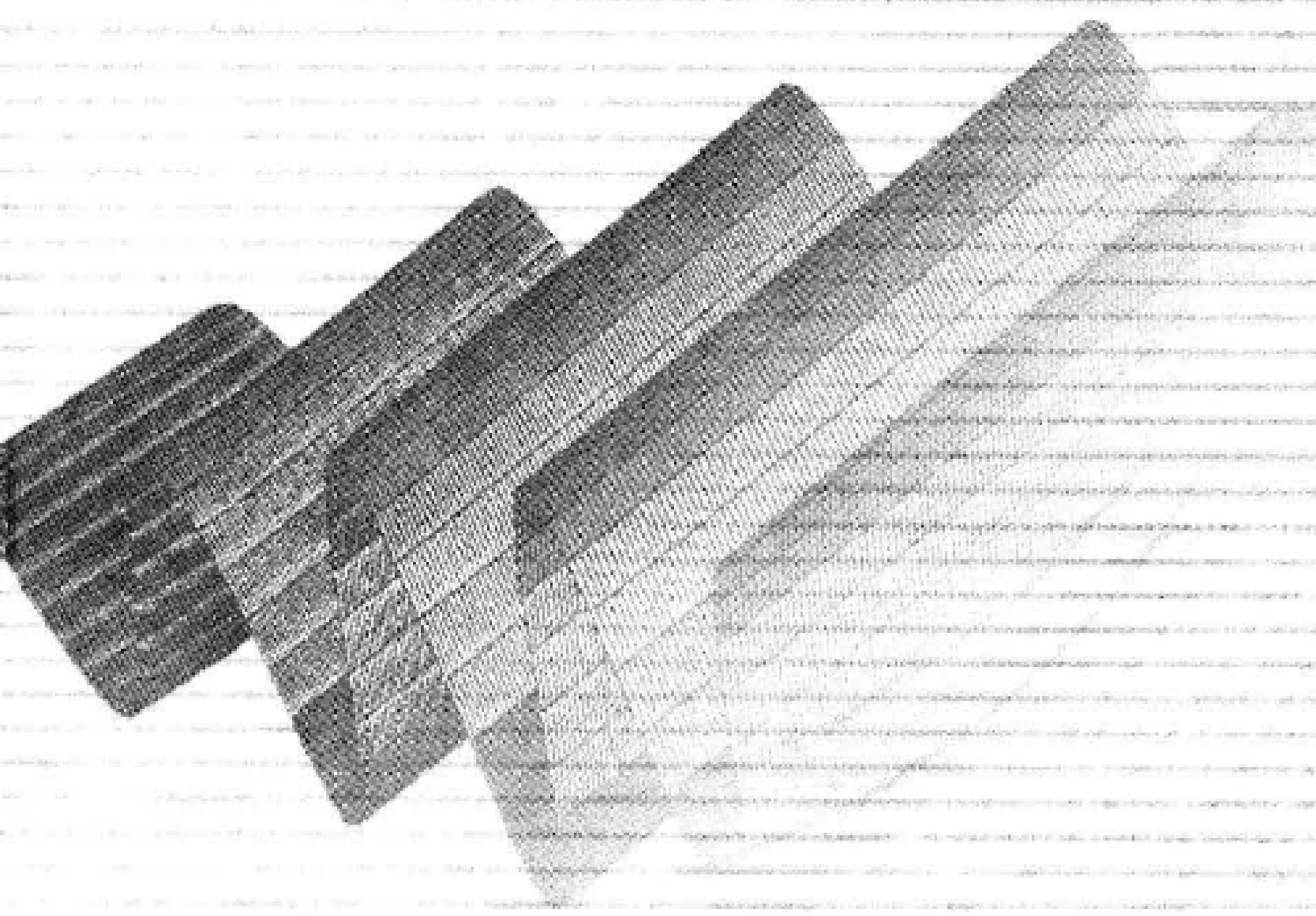


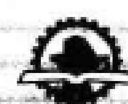
"十二五"规划教材

概率统计与SPSS应用

(第2版)

于义良 罗蕴玲 安建业 编著





西安交通大学出版社
XI'AN JIAOTONG UNIVERSITY PRESS

内容简介

本书是天津市普通高等学校本科教学质量与教学改革研究计划项目“基于应用型人才培养的大学数学综合改革与实践”(津教委高[2012]32号)的研究成果。其基本内容是依据国家非数学类专业数学教学指导委员会分委员会于2005年提出的关于“概率论与数理统计”课程教学基本要求确定的。本书将概率统计理论和世界公认的标准统计软件——SPSS软件相结合,基于SPSS软件介绍实际应用,易学易用。使用者在学习相关理论的基础上,可以轻松完成统计计算和分析,实现理论到实践的转化。

全书分为8章,内容包括随机事件及其概率、随机变量及其分布、随机向量及其分布、随机变量的数字特征;统计估值、统计检验、方差与协方差分析、相关与回归分析,以及20个演示实验。其特点是,内容可视化,计算软件化,方法现实化,技能突出,实用性强。

本书是天津市“十二五”规划教材,可作为高等学校非数学类本科专业学生的教材,也可作为应用统计工作者的参考书。

图书在版编目(CIP)数据

概率统计与SPSS应用/于义良,罗蕴玲,安建业编著.
—2版. —西安:西安交通大学出版社,2013.5
ISBN 978-7-5605-5200-2

Ⅰ.概… Ⅱ.①于…②罗…③安… Ⅲ.①概率统计-软件包-高等学校-教材 Ⅳ.0211-39

中国版本图书馆CIP数据核字(2013)第079899号

书　　名 概率统计与SPSS应用(第2版)
编　　著 于义良　罗蕴玲　安建业
责任编辑 任振国

出版发行 西安交通大学出版社
(西安市兴庆南路10号　邮政编码710049)
网　　址 http://www.xjtupress.com
电　　话 (029)82668357　82667874(发行中心)
(029)82668315　82669096(总编办)
传　　真 (029)82668280
印　　刷 西安明瑞印务有限公司

开　　本 727mm×960mm　1/16　**印张** 15.5　**字数** 286千字
版次印次 2013年5月第2版　2013年5月第1次印刷
书　　号 ISBN 978-7-5605-5200-2/O·426
定　　价 27.00元

读者购书、书店添货、如发现印装质量问题,请与本社发行中心联系、调换。
订购热线:(029)82665248　(029)82665249
投稿热线:(029)82664954
读者信箱:jdlgy@yahoo.cn

前言

随着科学技术的迅猛发展，数量分析已渗透到各个领域，数学的重要性已被整个社会所公认；由于计算机技术的广泛普及与提高，许多繁难的计算和抽象的推理已不再是高不可攀，数学的应用越来越深入；随着人类素质的不断提高，数学素质教育已成为全体公民的必修课，数学的普及越来越广泛。为适应新形势的发展和社会的需要，信息技术与学科课程整合已提到教育教学改革“重中之重”的地位，运用信息技术改造和优化传统学科内容是培养新世纪具有创新能力的高素质人才的必然要求。

天津商业大学“大学数学基础课程教学团队”是天津市级教学团队，经过多年的教学研究和实践，组织具有丰富教学经验的第一线教师，于 2009 年编著出版了《概率统计与 SPSS 应用》，经过三年多的教学实践，我们对书中的内容作了订正和调整，现奉献给大家。

《概率统计与 SPSS 应用》是天津市普通高等学校本科教学质量与教学改革研究计划项目——基于应用型人才培养的大学数学综合改革与实践（津教委高[2012]32 号）的成果，是天津市“十二五”规划教材。

这部教材力求体现如下特点：

第一，以实用为原则，“教、学、做”融为一体，内容体系整体优化，使读者实现由知识向能力的转化，比如分析、处理和解决复杂问题的能力，就业、创新和创业能力等。

第二，以实际为背景，概念阐述简明、通俗化，举例贴近生活，运用多媒体技术使内容直观化、图形化，使读者消除对数学的陌生感、抽象感、恐惧感，激活求知欲，增强学好数学、做好数学的信心，比如常见概率分布、抽样分布基本定理、置信区间、两类错误等。

第三，以计算机为工具，传统内容与信息技术应用有机融合，注重基本知识、基本思想、基本能力的培养，对繁、难、抽象的内容，充分利用当今极为流行的 SPSS 软件来实现，比如分布函数值的计算、点估计、区间估计、假设检验、方差分析、协方差分析、相关分析、回归分析等。

总之，本书融入软件，突出技能，实用性强。内容可视化，您不再被抽象而烦

恼；计算软件化，您不再被繁难而困扰；方法现实化，您不再被无用而厌学。

书中涉及的20个演示实验和SPSS数据文件均可在“天津市大学数学精品资源网”上下载，也可以和作者联系索取，E-mail:yuyil88@126.com。

主编于义良教授是天津市高等学校首届教学名师，曾到澳大利亚La Trobe大学学习考察，亲身经历了国外大学数学教育对学生能力、素质培养的实践，他们特别重视数学思想的熏陶和数学知识的应用，“做中学，学中悟，悟中醒，醒中行”做得非常出色。可喜的是本书恰好在这方面做了有益的尝试。

天津市教育委员会高教处、西安交通大学出版社对该项目的研究给予了热心的指导和支持，在此一并致以最诚挚的感谢。

我们期盼着本书能为广大读者带来学数学的轻松、做数学的快乐和效率。

编　者

2013.05

目　录

第4章　随机变量的数字特征

第5章　统计估值

第6章　统计检验

第7章　方差与协方差分析

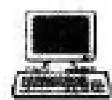
第 1 章　随机事件及其概率

生活在一个日新月异、千变万化的世界中,每个人时刻都要面对许多生活中碰到的问题。例如:"明天是雨天还是晴天,是否可以出去旅游";"明天的股市是上涨还是下跌,是买还是卖";"下个月某空调器的销售量是多少,如何组织货源";"今年夏季长江流域的降水量有多少,怎样组织抗洪";"在下届奥运会中我国体育健儿能拿多少金牌,如何争取金牌总数第一"等等,这些问题的发生与发展是受诸多因素的影响,因而这些问题的结果也是不确定的,不可预知的。但是,事实证明在许多不确定问题中隐藏着一种确定性的规律。也正因为如此,人类才在不断摸索和研究中使许多以前认为不可想象的问题得到解决,人类才取得了如此辉煌的进步。N. Weiner 说:"数学的伟大使命是在混沌中发现有序。"

本章的目的就是从基本问题出发,引出随机事件的概念,试图从最简单的随机现象(偶然现象)中去探求必然的规律。

1.1　随机事件

我们先做一个简单的试验:掷一颗质地均匀的骰子,观察出现的点数。

 **投掷骰子演示实验**

显然,这个试验具有如下特点:

(1) 在相同的条件下可以重复进行;

(2) 每次试验的可能结果不只一个(出现 1 点,出现 2 点,…,出现 6 点),而究竟出现哪个结果,在试验之前不能预言;

(3) 试验之前可以预知试验中一切可能的结果(六种结果),每次试验中出现且只出现可能结果中的一个。

我们把具有上述三个特点的试验称为**随机试验(random experiment)**,简称**试验(experiment)**,记为 E。本书中所谈的试验,均为随机试验。我们就是研究随机试验中这些可能结果出现的规律。

试验中的每个可能的结果称为**样本点(sample)**,用 ω 表示,全体样本点构成的空间称为**样本空间(sample space)**,用 Ω 表示。从集合论的观点看,样本空间就是针对某试验的所有可能结果构成的全集,而样本点就是构成样本空间的元素。样本空间 Ω 的子集称为**随机事件(random event)**,简称**事件(event)**,一般用大写字母 $\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\ldots$ 表示。我们称一事件在一次试验中出现(或发生)了,是指该次试验出现的结果(样本点)属于该事件(子集),否则称该事件没有出现(或发生)。易见 Ω 在每次试验中均要出现,故又称为**必然事件(certainty)**。Φ 在每次试验中均不出现,故称为**不可能事件(impossible event)**。

在上面掷骰子的试验中,我们设

ω_i 代表出现的点数为 $i(i=1, 2, \cdots,6)$,则 ω_i 为样本点,且记 $A_i=\{\omega_i\}$;

B=“出现的点数是 2 或者 3”;

C=“出现的点数大于 1 小于 5”;

D=“出现的点数是偶数”

则样本空间为:

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\};$$
$$B = \{\omega_2, \omega_3\};$$
$$C = \{\omega_2, \omega_3, \omega_4\};$$
$$D = \{\omega_2, \omega_4, \omega_6\}$$

显然 $A_i(i=1, 2, \cdots, 6)$,B, C, D 均为事件。而事件“出现的点数小于 7”是必定要出现的,它就是一个必然事件。实际上它包括了所有的样本点,即为 $\Omega=\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$。而事件“出现的点数大于 6”是不可能出现的,它就是一个不可能事件。实际上它不包含任何样本点,即为一个空集 Φ。

由集合和随机事件之间的关系,可得到如下的结论。

(1) $A\subset B$:A 是 B 的子集。它表示若事件 A 出现,则事件 B 一定出现。

(2)$A\cup B$(或 $A+B$):A 与 B 的并(或和)。它表示一个新的事件,即事件 A 和事件 B 至少有一个出现。同样事件 $\bigcup\limits_{i=1}^{n}A_i$ 表示 $A_1, A_2, \cdots, A_n$ 这 n 个事件至少有一个出现。

(3)$A\cap B$(或 AB):A 与 B 的交(积)。它表示一个新的事件,即事件 A 和事件 B 同时出现。同样事件 $\bigcap\limits_{i=1}^{n}A_i$ 表示 $A_1, A_2, \cdots, A_n$ 这 n 个事件同时出现。

(4)$A\cap B=\Phi$:它表示事件 A 和事件 B 不可能同时出现,我们称 A 与 B 为**互不相容**,简称**互斥(mutually exclusive events)**。

(5)$A\cap B=\Phi$ 且 $A\cup B=\Omega$:它表示事件 A 和事件 B 出现且只出现其中一个,我们称 A 与 B 为对立,并称 B 为 A(或 A 为 B)的**对立事件(complementary**

event),记为 $B=\overline{A}$(或 $A=\overline{B}$)。

(6)$A-B$:A 与 B 的差,它表示事件 A 出现而事件 B 不出现。显然 $A-B=A\overline{B}$,同时 $B-A=B\overline{A}$.

(7)$\overline{A\cup B}$:它是事件 A 和事件 B 至少出现一个的对立事件。显然它表示事件 A 和事件 B 都不出现,即$\overline{A\cup B}=\overline{A}\cap\overline{B}$。同样 $\overline{\bigcup\limits_{i=1}^{n} A_i}=\bigcap\limits_{i=1}^{n}\overline{A}_i$。

(8)$\overline{A\cap B}$:它是事件 A 和事件 B 同时出现的对立事件。显然它表示事件 A 和事件 B 至少有一个不出现,即$\overline{A\cap B}=\overline{A}\cup\overline{B}$。同样 $\overline{\bigcap\limits_{i=1}^{n} A_i}=\bigcup\limits_{i=1}^{n}\overline{A}_i$。

下面我们用图形(图 1.1.1)将上面的结论直观地显示出来,见图 1.1.1。

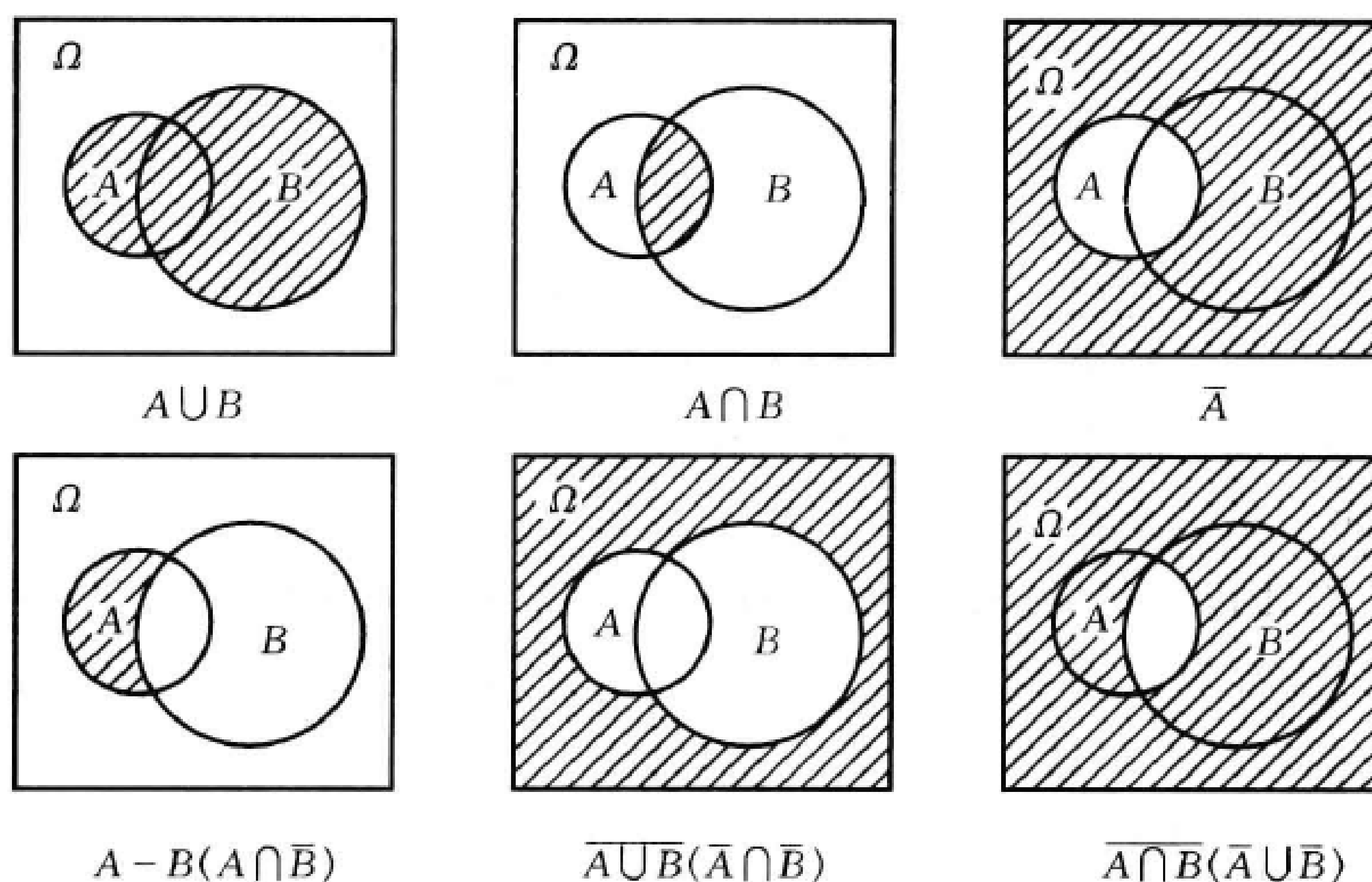


图 1.1.1

从上面的图示中,还可以得到以下一些关系:

$A\cap B\subset A$;$A-B\subset A$,且 $A-B=A-AB=A\overline{B}$,$AB\subset A$;$A\overline{B}$,$\overline{A}B$ 与 AB 两两互斥,$A=A\overline{B}\cup AB$,$A\cup B=A\overline{B}\cup B\overline{A}\cup AB=A\overline{B}\cup B=B\overline{A}\cup A=A\cup(B-A)$;$\overline{A\cap B}\neq\overline{A}\cap\overline{B}$;$\overline{A\cup B}\neq\overline{A}\cup\overline{B}$,…

结合前面的试验以及字母 $\omega_i(i=1,2,\cdots,6)$,B,C,D 表示的意义,我们有

$B\subset C$;　　　　$C\cup D=\{\omega_2,\omega_3,\omega_4,\omega_6\}$;

$C\cap D=\{\omega_2,\omega_4\}$;$\overline{B}=\{\omega_1,\omega_4,\omega_5,\omega_6\}$;

$\overline{C\cup D}=\{\omega_1,\omega_5\}$;$\overline{C\cap D}=\{\omega_1,\omega_3,\omega_5,\omega_6\}$;

$C-D=\{\omega_3\}$;　　$D-C=\{\omega_6\}$

大家还可以验证其它的一些等式或不等式。

以上一些表示法，为我们今后处理某些复杂事件带来很大方便。

练习 1.1

1. 将一枚均匀的硬币抛两次，事件 A，B，C 分别表示“第一次出现正面”，“两次出现同一面”，“至少有一次出现正面”。试写出样本空间及事件 A，B，C 中的样本点。

2. 在掷两颗骰子的试验中，事件 A，B，C，D 分别表示“点数之和为偶数”，“点数之和小于 5”，“点数相等”，“至少有一颗骰子的点数为 3”。试写出样本空间及事件 AB，$A+B$，$\overline{A}C$，BC，$A-B-C-D$ 中的样本点。

3. 以 A，B，C 分别表示某城市居民订阅日报、晚报和体育报。试用 A，B，C 表示以下事件：

(1)只订阅日报； (2)只订日报和晚报；

(3)只订一种报； (4)正好订两种报；

(5)至少订阅一种报； (6)不订阅任何报；

(7)至多订阅一种报； (8)三种报纸都订阅；

(9)三种报纸不全订阅。

4. 甲、乙、丙三人各射击一次，事件 A_1，A_2，A_3 分别表示甲、乙、丙射中。试说明下列事件所表示的结果：$\overline{A_2}$，A_2+A_3，$\overline{A_1A_2}$，$\overline{A_1+A_2}$，$A_1A_2\overline{A_3}$，$A_1A_2+A_2A_3+A_1A_3$。

5. 设事件 A，B，C 满足 $ABC\neq\Phi$，试把下列事件表示为一些互不相容的事件的和：$A+B+C$，$AB+C$，$B-AC$。

6. 若事件 A，B，C 满足 $A+C=B+C$，试问 $A=B$ 是否成立？举例说明。

7. 对于事件 A,B,C，试问 $A-(B-C)=(A-B)+C$ 是否成立？举例说明。

1.2 事件的概率

下面再回到原来的试验：掷一颗质地均匀的骰子，观察出现的点数。

显然，这个试验具有两个特点：

(1) 所有可能的试验结果是有限个(有限性)；

(2) 每个可能结果在一次试验中出现的可能性相同(等可能性)。

我们把具有这两个特点的试验称为**古典概型(classical probability)**。

我们知道在每次试验中，出现且只出现样本空间中的一个样本点。所谓事件 A 出现，就是指试验出现事件 A 中包含的样本点。因此，在古典概型中由于样本

点出现的等可能性，事件 A 出现的可能性大小就可以用事件 A 中包含的样本点的个数占样本点总数的比例来度量。

定义 1.2.1　在古典概型中，设样本空间 Ω 中包含有 n 个样本点，则对任意事件 A，若 A 中含有 k 个样本点，那么事件 A 的概率 $P(A)$ 定义为

$$P(A)=\frac{\text{事件 } A \text{ 中包含的样本点数}}{\text{样本空间 } \Omega \text{ 中样本点总数}}=\frac{k}{n}$$

在掷骰子的试验中，显然有

$P(A_i)=\frac{1}{6},\ i=1,\ 2,\ 3,\ 4,\ 5,\ 6$；　　$P(B)=\frac{2}{6}$；

$P(C)=\frac{3}{6}$；$P(D)=\frac{3}{6}$；　　$P(C\cup D)=\frac{4}{6}$；

$P(C\cap D)=\frac{2}{6}$；$P(\overline{B})=\frac{4}{6}$；　　$P(\Omega)=\frac{6}{6}=1$；

$P(\Phi)=\frac{0}{6}=0$

在计算古典概型概率中，往往要用到中学时的排列与组合的知识。

例 1.2.1　箱中装有 10 件产品，其中 1 件是次品，在 9 件合格品中有 6 件是一等品，3 件二等品。现从箱中任取 3 件，试求：① 取得 3 件产品都是一等品的概率；② 取得 3 件产品中有 1 件是一等品，2 件是二等品的概率；③ 取得 3 件产品中至少有 2 件是一等品的概率。

解　由于试验中是任取 3 件，所以这个试验是古典概型。每个样本点就是从 10 件中任取 3 件产品构成的集合，与顺序无关，故样本空间中样本点的总数为 C_{10}^3。

① 设 A＝“取得 3 件产品都是一等品”，那么 A 中的样本点个数为 C_6^3，所以

$$P(A)=\frac{C_6^3}{C_{10}^3}=\frac{1}{6}$$

② 设 B＝“取得 3 件产品中有 1 件是一等品，2 件是二等品”，那么 B 中的样本点个数为 $C_6^1\cdot C_3^2$（这里用到了乘法原理），所以

$$P(B)=\frac{C_6^1\cdot C_3^2}{C_{10}^3}=\frac{3}{20}$$

③ 设 C＝“取得 3 件产品中至少有 2 件是一等品”，那么事件 C 显然是由“有 2 件一等品，1 件非一等品”和“3 件都是一等品”两个事件构成，所以 C 中的样本点个数为 $C_6^2\cdot C_4^1+C_6^3$（这里又用到了加法原理），那么

$$P(C)=\frac{C_6^2C_4^1+C_6^3}{C_{10}^3}=\frac{2}{3}$$

生活中还有很多问题并不具有古典概型的两个特点，例如投掷的是一颗不均

匀的骰子，那么样本空间中的 6 个样本点出现的可能性就不相等了，因此计算其中的一些事件的概率时就不能再用古典概型的公式了。如何解决这类问题呢？我们最常用也最直接的办法就是反复掷这颗骰子（例如做了 n 次试验），记下所关心的事件 A 在这 n 次试验中出现的次数（例如出现了 μ 次），比值 $\frac{\mu}{n}$（即事件 A 在这 n 次试验中出现的频数）随着试验次数 n 的增大而越来越接近某个数值 p（这被称为频率的稳定性），那么我们就称数值 p 为事件 A 的概率，即 $P(A)=p$，有的书上称此为事件概率的统计定义. 但是数值 p 是个理论上的数，在有限次的试验中很难求得。一般就用 $\frac{\mu}{n}$ 作为 $P(A)$ 的近似值，即 $P(A)\approx\frac{\mu}{n}$。

上面我们介绍了两种计算随机事件概率的方法。实际上由于问题的不同以及处理问题的角度不同，还有许多计算随机事件概率的方法，但是不管用什么方法计算概率，它们都要求具有下面三个基本性质：

(1) **非负性(nonnegativity)**：$0\leqslant P(A)\leqslant 1$；

(2) **规范性(normativity)**：$P(\Omega)=1, P(\Phi)=0$；

(3) **可列可加性(additivity)**：若 $A_i\ (i=1,\ 2,\ \cdots)$ 是两两互不相容的事件（即 $A_iA_j=\Phi,\ i\neq j$），则 $P(\bigcup\limits_{i=1}^{\infty} A_i = \sum\limits_{i=1}^{\infty} P(A_i)$ 。

以上三个性质也被称为**概率的公理化定义(axiomatized definition)**。即对任意事件 A，定义实函数 $P(A)$，如果此实函数同时满足上述三个性质（或称三条公理），那么就称 $P(A)$ 为 A 的概率。

由此公理化定义可以推出以下一些性质：

(1)（有限可加性）若 $A_1,\ A_2,\ \cdots,\ A_n$ 两两互不相容，则 $P(\bigcup\limits_{i=1}^{n} A_i) = \sum\limits_{i=1}^{n} P(A_i)$ ；

(2)（逆算公式）$P(\overline{A})=1-P(A)$；

(3)（减法公式）若 $A\supset B$，则 $P(A-B)=P(A)-P(B)$；

(4)（一般减法公式）$P(A-B)=P(A\overline{B})=P(A)-P(AB)$；

(5)（一般加法公式）$p(A\cup B)=P(A)+P(B)-P(AB)$；

(6)（次可加性）$P(A\cup B)\leqslant P(A)+P(B)$；

(7)（单调不减性）若 $A\supset B$，则 $P(A)\geqslant P(B)$。

证 我们只证(2)，(3)，(5)，其余几条很容易推得（在证明性质 2 时用到了性质 1）。

(2) 因为 $\Omega=A\cup\overline{A}$，所以 $1=P(\Omega)=P(A\cup\overline{A}=P(A)+P(\overline{A})$，故 $P(\overline{A})=1-P(A)$。

(3) 因为 $A\supset B$,所以 $A=(A-B)\cup B$,又因 $(A-B)\cap B=\Phi$,于是得 $P(A)=P[(A-B)\cup B]=P(A-B)+P(B)$,故 $P(A-B)=P(A)-P(B)$。

(5) 因为 $A\cup B=A\cup(B-AB)$,而 $A\cap(B-AB)=\Phi$,所以

$$\begin{aligned}P(A\cup B)&=P[A\cup(B-AB)=P(A)+P(B-AB)\\&=P(A)+P(B)-P(AB)\end{aligned}$$

通过上面的证明过程可以看出,善于把一个事件分解成互斥事件是一个常用的技巧。

例 1.2.2　假设 A 出现的概率为 0.6,A 与 B 都出现的概率为 0.1,A 与 B 都不出现的概率为 0.15,求:① A 出现但是 B 不出现的概率;② A 与 B 至少出现一个的概率。

解　依题意 $P(A)=0.6$, $P(AB)=0.1$, $P(\overline{A}\,\overline{B})=0.15$,于是

① $P(A\overline{B})=P(A-B)=P(A-AB)=P(A)-P(AB)=0.5$;

② $P(A\cup B)=1-P(\overline{A\cup B})=1-P(\overline{A}\,\overline{B})=0.85$。

练习 1.2

1. 设 $P(A)=\dfrac{1}{3}$,$P(B)=\dfrac{1}{2}$,试就以下三种情况分别求 $P(B\overline{A})$:

(1)$AB=\Phi$;　　(2)$A\subset B$;(3)$P(AB)=\dfrac{1}{8}$。

2. 已知 $P(A)=P(B)=P(C)=\dfrac{1}{4}$, $P(AC)=P(BC)=\dfrac{1}{16}$,$P(AB)=0$,求事件 A, B, C 全不发生的概率。

3. 每个路口有红、绿、黄三色指示灯,假设各色灯的开闭是等可能的。一个人骑车经过三个路口,试求下列事件的概率:A=“三个都是红灯”=“全红”; B=“全绿”; C=“全黄”;D= “无红”; E=“无绿”; F=“三次颜色相同”;G= “颜色全不相同”;H= “颜色不全相同”。

4. 设一批产品共 100 件,其中 98 件正品,2 件次品,从中任意抽取 3 件(分三种情况:一次拿 3 件;每次拿 1 件,取后放回拿 3 次;每次拿 1 件,取后不放回拿 3 次),试求

(1) 取出的 3 件中恰有 1 件是次品的概率;

(2) 取出的 3 件中至少有 1 件是次品的概率。

5. 从 0, 1, 2, …, 9 中任意选出 3 个不同的数字,试求下列事件的概率:A_1={三个数字中不含 0 与 5}, A_2={三个数字中不含 0 或 5}。

6. 从 0, 1, 2, …,9 中任意选出 4 个不同的数字,计算它们能组成一个 4 位偶

数的概率。

7. 一个宿舍中住有 6 位同学，计算下列事件的概率：

(1)6 人中至少有 1 人生日在 10 月份的概率；

(2)6 人中恰有 4 人生日在 10 月份的概率；

(3)6 人中恰有 4 人生日在同一月份的概率。

8. 从一副扑克牌(52 张)任取 3 张(不重复)，计算取出的 3 张牌中至少有 2 张花色相同的概率。

1.3 条件概率与独立性及其应用

1. 条件概率

我们来看一个试验：箱中装有 10 件产品，其中 2 件次品，8 件正品，先后有 2 人买此产品，每人 1 件，甲先乙后。

(1)已知甲买走 1 件正品(A)，而乙在该箱中买的又是正品(B)；

(2)已知甲买走 1 件正品，乙要求另开 1 箱，且乙买的也是正品；

(3)甲急急忙忙买走 1 件产品，乙在该箱中买的是正品。

以上三种情况的结果只有一个：乙买走了正品(B)，但由于前提条件不同，因此 B 发生的概率也就不同。

下面我们先讨论第一种情况。

用 $P(B|A)$ 来表示该事件的概率。由于乙在买产品时箱中有 9 件产品，其中 2 件次品，7 件正品，于是由古典概型的公式得 $P(B|A)=\frac{7}{9}$。

我们再进一步讨论这个问题。由古典概型的公式易得 $P(A)=\frac{8}{10}$，$P(AB)=\frac{C_8^1 \cdot C_7^1}{C_{10}^1 \cdot C_9^1}=\frac{8\times 7}{10\times 9}$，那么 $\frac{P(AB)}{P(A)}=\frac{7}{9}$，此时有

$$P(B \mid A) = \frac{P(AB)}{P(A)}$$

事实上，这个公式有普遍的意义。因为讨论事件($B|A$)时是在事件 A 包含的样本点内考虑事件 AB，即 $P(B|A)$ 是在缩小了的样本空间 A 中讨论的。

定义 1.3.1 对于两个事件 A 与 B，如果 $P(A)>0$，称

$$P(B|A)=\frac{P(AB)}{P(A)}$$

为事件 A 先出现的条件下事件 B 后出现的**条件概率(conditional probability)**。

当 $A=\Omega$ 时，条件概率 $P(B|\Omega)=P(B)$。这也正反映了条件概率与无条件概率之间的区别与联系，另外 $P(AB)$ 与 $P(B|A)$ 不同，$P(AB)$ 是在 Ω 中考虑 AB 的（如图1.3.1）。

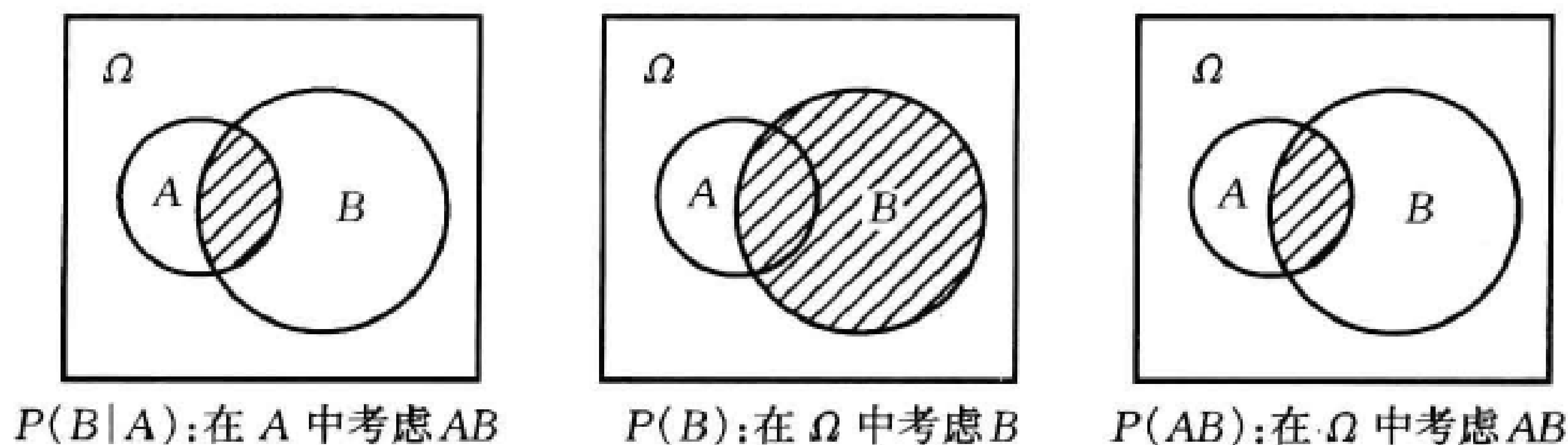


图1.3.1

在计算条件概率时，一般有两种方法：

(1)公式法，$P(B|A)=\dfrac{P(AB)}{P(A)}$，$P(A)>0$。

(2)由$(B|A)$的实际意义，按古典概型公式直接计算。

由条件概率的定义很容易得到下面的公式：

当 $P(A)>0$ 时，$P(AB)=P(A)\cdot P(B|A)$；

当 $P(B)>0$ 时，$P(AB)=P(B)\cdot P(A|B)$；

当 $P(AB)>0$ 时，$P(ABC)=P(A)\cdot P(B|A)\cdot P(C|AB)$

以上公式称为**乘法公式(multiplication formula)**。并在相应条件成立情况下，此结论可推广到任意有限个。即当 $P(A_1A_2\cdots A_{n-1})>0$ 时，有

$$P(A_1A_2\cdots A_n)=P(A_1)\cdot P(A_2|A_1)\cdots P(A_n|A_1A_2\cdots A_{n-1})$$

下面我们先讨论第二种情况。

由于乙要求另开一箱，此时箱中仍有10件产品，其中2件次品，8件正品，所以 $P(B|A)=\dfrac{8}{10}$。我们发现，这种情况下乙买到正品(B)的概率并不受甲是否买到正品(A)的影响，即 $$P(B|A)=P(B)=\frac{8}{10}。$$

由乘法公式即得　$P(AB)=P(A)\cdot P(B)$。

2. 事件的独立性

定义1.3.2　如果两个事件 A 与 B 满足等式

$$P(AB)=P(A)\cdot P(B)$$

称事件 A 与 B 是**相互独立的(mutually independent events)**，简称 **A 与 B 独立**

(independent)。

从直观上讲，A 与 B 独立就是其中任何一个事件出现的概率不受另一个事件出现与否的影响。

推论 1.3.1 A 与 B 为两个事件，当 $P(B)>0$ 时，A 与 B 独立的充分必要条件是

$$P(A|B)=P(A)$$

当 $P(A)>0$ 时，A 与 B 独立的充分必要条件是

$$P(B|A)=P(B)$$

推论 1.3.2 设 A 与 B 为两个事件，则下列四对事件：A 与 B；$\overline{A}$与 B；A 与$\overline{B}$；$\overline{A}$与$\overline{B}$中，只要有一对事件独立，则其余三对也独立。

推论 1.3.1 的证明很容易从乘法公式中推出。

下面只证推论 1.3.2 中的一种情况，其它情况可类似推出。

证 （不妨设 A 与 B 独立，我们证$\overline{A}$与 B 也独立）

因为$\overline{A}B=B-AB$，且 $AB\subset B$，所以

$$\begin{aligned} P(\overline{A}B) &= P(B-AB) = P(B)-P(AB) \\ &= P(B)-P(A)\cdot P(B) \\ &= P(B)\cdot(1-P(A)) \\ &= P(\overline{A})\cdot P(B) \end{aligned}$$

故$\overline{A}$与 B 独立。

其实推论 1.3.2 的证明也可以这样思考：首先由 A 与 B 独立，证$\overline{A}$与 B 也独立；类似可证 A 与$\overline{B}$也独立；同理由$\overline{A}$与 B（或 A 与$\overline{B}$）独立，可证$\overline{A}$与$\overline{B}$独立；进一步由$\overline{A}$与$\overline{B}$独立，可证 A 与 B 独立。

判断事件的独立性一般有两种方法：

(1) 由定义判断，是否满足公式；

(2) 由问题的性质从直观上去判断。

以上关于独立性的概念可推广到任意有限个。

定义 1.3.3 设 A_1, A_2, …, A_n 为 n 个事件，如果对任何正整数 $m(2\leqslant m\leqslant n)$以及 $1\leqslant i_1\leqslant i_2\leqslant\cdots\leqslant i_m\leqslant n$，都有

$$P(A_{i_1}A_{i_2}\cdots A_{i_m})=P(A_{i_1})\cdot P(A_{i_2})\cdots P(A_{i_m})$$

称事件 A_1, A_2, …, A_n 是相互独立。

如果上式仅对 $m=2$ 成立，则称事件 A_1, A_2, …, A_n 两两独立。显然 n 个事件($n>2$)的相互独立性和两两独立性不同，前者比后者要求条件强。

从直观上讲，A_1, A_2, …, A_n 相互独立，就是指它们中任何一个事件出现的概率不受其余某一个或几个事件出现与否的影响。这 n 个事件中的任何部分事件都

是相互独立的。$A_1, A_2, \cdots, A_n$ 两两独立,就是指这 n 个事件中的任何两个事件是相互独立的。显然,由事件的相互独立可推出事件的两两独立,但反之不成立。

例 1.3.1　设甲、乙、丙三人在同一时间内破译某个密码,而甲、乙、丙三人单独能译出的概率分别为 0.8,0.7 和 0.6,求:①密码能译出的概率;②最多只有一人能译出的概率。

解　设 A=“甲译出密码”,B,C 分别表示乙、丙能译出密码,D=“密码被破译”,F=“最多只有一人译出”。依实际情况,A,B,C 相互独立,则

①
$$\begin{aligned}P(D)&=P(A+B+C)\\&=P(A)+P(B)+P(C)-P(A)P(B)-P(A)P(C)-P(B)P(C)+\\&\quad P(A)P(B)P(C)\\&=0.8+0.7+0.6-0.8\times0.7-0.8\times0.6-0.7\times0.6+\\&\quad 0.8\times0.7\times0.6\\&=0.976\end{aligned}$$

当然还可以这样做

$$\begin{aligned}P(\overline{D})&=P(\overline{A+B+C})=P(\overline{A}\,\overline{B}\,\overline{C})\\&=P(\overline{A})\cdot P(\overline{B})\cdot P(\overline{C})\\&=(1-0.8)\times(1-0.7)\times(1-0.6)\\&=0.024\end{aligned}$$

故

$$P(D)=1-P(\overline{D})=1-0.024=0.976$$

② 因为 $F=\overline{ABC}+A\,\overline{BC}+\overline{AB}\,\overline{C}+\overline{A}\,B\,\overline{C}$,而$\overline{ABC}$,$A\,\overline{BC}$,$\overline{AB}\,\overline{C}$,$\overline{AB}C$ 是两两互斥,所以

$$\begin{aligned}P(F)&=P(\overline{A}\,\overline{B}\,\overline{C}+A\,\overline{B}\,\overline{C}+\overline{A}B\,\overline{C}+\overline{A}\,\overline{B}C)\\&=P(\overline{A}\,\overline{B}\,\overline{C})+P(A\,\overline{B}\,\overline{C})+P(\overline{A}B\,\overline{C})+P(\overline{A}\,\overline{B}C)\\&=P(\overline{A})\cdot P(\overline{B})\cdot P(\overline{C})+P(A)\cdot P(\overline{B})\cdot P(\overline{C})+\\&\quad P(\overline{A})\cdot P(B)\cdot P(\overline{C})+P(\overline{A})\cdot P(\overline{B})\cdot P(C)\\&=0.2\times0.3\times0.4+0.8\times0.3\times0.4+0.2\times0.7\times0.4+\\&\quad 0.2\times0.3\times0.6\\&=0.212\end{aligned}$$

3. 应　用

我们再回去讨论前面试验中的第三种情况。

由于不知道甲买走的是正品还是次品,这就增加了问题的难度。为此我们从实际情况出发,全面考虑问题。

显然，$B=BA+B\overline{A}$，其中 $A\overline{A}=\Phi$，$A+\overline{A}=\Omega$，那么

$$\begin{aligned}P(B) &= P(BA+B\overline{A}) \\ &= P(BA)+P(B\overline{A}) \\ &= P(A)P(B\mid A)+P(\overline{A})P(B\mid\overline{A}) \\ &= \frac{8}{10}\times\frac{7}{9}+\frac{2}{10}\times\frac{8}{9} \\ &= \frac{36}{45}\end{aligned}$$

按照上面处理问题的思路和方法，可以得出下面两个很重要的公式，它们对于计算较为复杂的事件概率是很有用的。

全概率公式 设 Ω 为随机试验 E 的样本空间，事件组 $A_1, A_2, \cdots, A_n$ 满足

(1) $A_iA_j=\Phi,\ i\neq j$；

(2) $\bigcup\limits_{i=1}^{n}A_i=\Omega$，$P(A_i)>0,\ i=1,2,\cdots,n$。

则对任一事件 B，有

$$P(B)=\sum_{i=1}^{n}P(A_i)\cdot P(B\mid A_i)。$$

这个公式的直观意义很清楚：把复杂事件（直接计算其概率比较难）B 分解成两两互斥的若干个简单事件（可以直接计算其概率）的并（见图 1.3.2）。

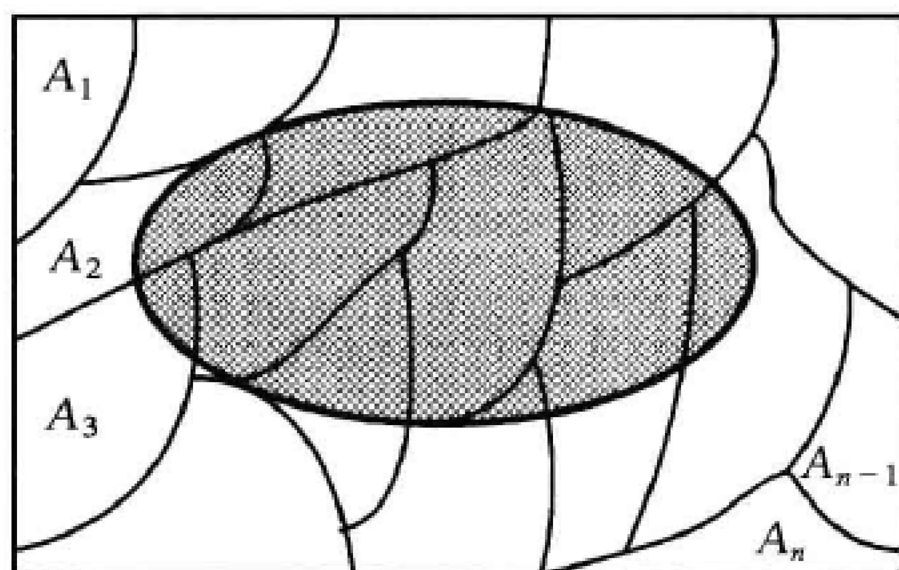


图 1.3.2

其中 B 为黑色圈内样本点的集合，且 $B=$"B 中互不相交的子集的并集"。

证 $P(B)=P(B\Omega)=P[B(A_1+A_2+\cdots+A_n)]$

$\qquad=P(BA_1)+P(BA_2)+\cdots+P(BA_n)$

由于 $A_1, A_2, \cdots, A_n$ 两两互斥，所以 $BA_1, BA_2, \cdots, BA_n$ 也两两互斥，再由乘法公式可得

$$P(B)=\sum_{i=1}^{n}P(BA_i)=\sum_{i=1}^{n}P(A_i)\cdot P(B\mid A_i)。$$

在上面的试验中，如果再增加一问：已知乙买到的是正品，那么甲买到正品的概率是多少？这显然是求条件概率 $P(A|B)$，由条件概率的定义及乘法公式得

$$\begin{aligned}P(A \mid B) &= \frac{P(AB)}{P(B)} \\ &= \frac{P(A)\cdot P(B \mid A)}{P(A)\cdot P(B \mid A)+P(\overline{A})\cdot P(B \mid \overline{A})} \\ &= \frac{\frac{8}{10}\times\frac{7}{9}}{\frac{8}{10}\times\frac{7}{9}+\frac{2}{10}\times\frac{8}{9}} \\ &= \frac{7}{9}\end{aligned}$$

把上面的结论推广一下，就得到 Bayes 公式。

Bayes 公式 设 Ω 为随机试验 E 的样本空间，事件组 $A_1, A_2, \cdots, A_n$ 满足

(1) $A_iA_j=\Phi, i\neq j$ ；

(2) $\bigcup\limits_{i=1}^{n}A_i=\Omega, P(A_i)>0, i=1, 2, \cdots, n$。

则对任一概率不为零的事件 B，有

$$P(A_j \mid B)=\frac{P(A_j)\cdot P(B \mid A_j)}{\sum\limits_{i=1}^{n}P(A_i)\cdot P(B \mid A_i)}$$

在用上面全概率公式和 Bayes 公式计算时，关键是找到与事件 B 有关且满足两个条件的事件组 $A_1, A_2, \cdots, A_n$。事件组 $A_1, A_2, \cdots, A_n$ 一般是导致事件 B 出现的因素。

例 1.3.2 每箱产品有 10 件，其中的次品数从 0 到 2 是等可能的。开箱试验时，从中一次抽取 2 件（不重复），如果发现有次品，则拒收该箱产品。试计算：① 一箱产品通过验收的概率；② 已知该箱产品通过验收，则该箱中有 2 个次品的概率。

解 设 A_i =“箱内有 i 件次品”，$i=0, 1, 2$；B =“该箱产品通过验收”。

显然，A_0, A_1, A_2 满足 $A_iA_j=\Phi(i\neq j)$；$\bigcup\limits_{i=0}^{2}A_i=\Omega$，而且

$$P(A_i)=\frac{1}{3}(i=0, 1, 2), \qquad P(B|A_0)=1,$$

$$P(B|A_1)=\frac{C_9^2}{C_{10}^2}, \qquad P(B|A_2)=\frac{C_8^2}{C_{10}^2}$$

① 由全概率公式，有

$$P(B)=P(A_0)P(B \mid A_0)+P(A_1)P(B \mid A_1)+P(A_2)P(B \mid A_2)$$

$$=\frac{1}{3}\times 1+\frac{1}{3}\times\frac{C_9^2}{C_{10}^2}+\frac{1}{3}\times\frac{C_8^2}{C_{10}^2}=0.807$$

② 由 Bayes 公式，有

$$P(A_2|B)=\frac{P(A_2)\cdot P(B|A_2)}{P(B)}=\frac{\frac{1}{3}\times\frac{C_8^2}{C_{10}^2}}{0.807}=0.257$$

下面看一个较复杂一点的关于独立性的试验。现有五箱产品，每箱中均有 10 件产品，其中 3 件次品，7 件正品。现从每箱中均任取 1 件产品，共取得 5 件产品，我们研究这 5 件产品中正品数的情况。

这个试验是由 5 次基本试验(从每箱中任取 1 件产品为一次基本试验)构成，它显然有以下特点：

(1) 每次基本试验中只有两个结果：正品(A)，次品($\overline{A}$)；

(2) 每次基本试验中每个结果出现的概率不变：$P(A)=\frac{7}{10}$，$P(\overline{A})=\frac{3}{10}$；

(3) 基本试验之间相互独立；

(4) 在相同条件下，试验可以重复进行。

具有以上四个特点的试验称为独立重复试验，也称为**贝努里(Bernoulli)试验**。若该试验由 n 次基本事件构成，就称为 ***n* 重贝努里试验**。上面的试验就是一个 5 重贝努里试验。这种概率模型在理论和实践方面具有重要的意义。

这 5 件产品中正品数的情况如何呢？我们研究其中一种情况。令 B_3 ="5 件产品中恰有 3 件正品"，A_i ="第 i 个箱中取出的是正品"，$i=1, 2, 3, 4, 5$；显然

$$B=A_1A_2A_3\overline{A}_4\overline{A}_5+A_1A_2\overline{A}_3\overline{A}_4A_5+\cdots+A_1\overline{A}_2A_3\overline{A}_4\overline{A}_5+\cdots+\overline{A}_1\overline{A}_2A_3A_4A_5$$

即 B 分解为若干互不相容事件的并。这并中到底有多少事件呢？它的个数恰好为从 5 个元素中任取出 3 个元素的组合数 C_5^3。由于独立性及 $P(A)$，$P(\overline{A})$的不变，那么这 C_5^3 个事件的概率相等，且都等于$\left(\frac{7}{10}\right)^3\cdot\left(\frac{3}{10}\right)^2$。所以

$$P(B_3)=C_5^3\left(\frac{7}{10}\right)^3\left(\frac{3}{10}\right)^2$$

根据同样的思想，令 B_k ="5 件产品中恰有 k 件正品"，$k=0, 1, 2, 3, 4, 5$，那么

$$P(B_k)=C_5^k\left(\frac{7}{10}\right)^k\left(\frac{3}{10}\right)^{5-k}\qquad k=0, 1, 2, 3, 4, 5$$

将上面的公式推广一下，就得到重要的贝努里公式。

贝努里公式 在 n 重贝努里试验中，如果事件 A 在每次试验中出现的概率为

p，令 B_k＝“在 n 次试验中事件 A 恰好出现 k 次”，那么

$$P(B_k) = C_n^k p^k (1-p)^{n-k} \qquad k = 0, 1, 2, \cdots, n$$

这个公式在下一章还要用到。

例 1.3.3　同时掷四颗均匀的骰子，试计算：① 恰有一颗是 6 点的概率；② 至少有一颗是 6 点的概率。

解　这是一个 4 重贝努里试验，掷每一颗骰子就是一个基本试验。每一颗骰子“出现 6 点”(A)的概率均为$\frac{1}{6}$，A 的对立事件“不出现 6 点”($\overline{A}$)的概率均为$\frac{5}{6}$，那么

① $P(B_1)=C_4^1\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^3$；

② B＝“至少有一颗是 6 点”，那么 B 的对立事件为 B_0，所以

$$P(B) = 1 - P(B_0) = 1 - C_4^0\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^4$$

当然也可以这样计算(不如上面的方法简单)

$$\begin{aligned}P(B) &= P(B_1 + B_2 + B_3 + B_4) = \sum_{i=1}^{4} P(B_i) \\ &= C_4^1\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^3 + C_4^2\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^2 + C_4^3\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^1 + C_4^4\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^0\end{aligned}$$

下面再看一道例题，要分清不同试验的特点。

例 1.3.4　设一箱中装有 a 件正品和 b 件次品，试求下列事件的概率：

① 有放回地取 3 次，每次 1 件，取得 3 产品依次为次品、正品、次品；

② 有放回地取 3 次，每次 1 件，取得 2 件次品 1 件正品；

③ 无放回地取 3 次，每次 1 件，取得 3 件产品依次为次品、正品、次品；

④ 无放回地取 3 次，每次 1 件，取得 2 件次品 1 件正品；

⑤ 一次拿 3 件产品取得 2 件次品 1 件正品。

解　(这里只给出简略解答，读者试作详细分析)

① $p=\frac{b}{a+b}\cdot\frac{a}{a+b}\cdot\frac{b}{a+b}=\frac{ab^2}{(a+b)^3}$；

② $p=C_3^2\left(\frac{b}{a+b}\right)^2\left(\frac{a}{a+b}\right)$；

③ $p=\frac{b}{a+b}\cdot\frac{a}{a+b-1}\cdot\frac{b-1}{a+b-2}$；

④ $p=\frac{b}{a+b}\cdot\frac{b-1}{a+b-1}\cdot\frac{a}{a+b-2}+\frac{b}{a+b}\cdot\frac{a}{a+b-1}\cdot\frac{b-1}{a+b-2}+\frac{a}{a+b}\cdot$

$$\frac{b}{a+b-1}\cdot\frac{b-1}{a+b-2}$$
$$=\frac{3ab(b-1)}{(a+b)(a+b-1)(a+b-2)};$$

⑤ $P=\frac{C_a^1C_b^2}{C_{a+b}^3}=\frac{3ab(b-1)}{(a+b)(a+b-1)(a+b-2)}$。

注意 最后两个事件(4)与(5)的概率相等。

练习 1.3

1. 假设一批产品中一、二、三等品各占 60%,30%,10%,从中任取一件,结果不是三等品,求取到是一等品的概率。

2. 设 10 件产品中有 4 件不合格品,从中任取 2 件,已知所取 2 件产品中有 1 件不合格品,求另一件也是不合格品的概率。

3. 为了防止意外,在矿内同时装有两种报警系统(I)和(II)。两种报警系统单独使用时,系统(I)和(II)有效的概率分别为 0.92 和 0.93,在系统(I)失灵的条件下,系统(II)仍有效的概率为 0.85,求

(1) 两种报警系统(I)和(II)都有效的概率;

(2) 系统(II)失灵而系统(I)有效的概率;

(3) 在系统(II)失灵的条件下;系统(I)仍有效的概率。

4. 设 $0<P(A)<1$,证明:事件 A 与 B 独立的充要条件是

$$P(B\mid A)=P(B\mid\overline{A})$$

5. 设事件 A 与 B 相互独立,两个事件只有 A 发生的概率与只有 B 发生的概率都是$\frac{1}{4}$,求 $P(A)$和 $P(B)$。

6. 证明:若 $P(A)>0$,$P(B)>0$,则有

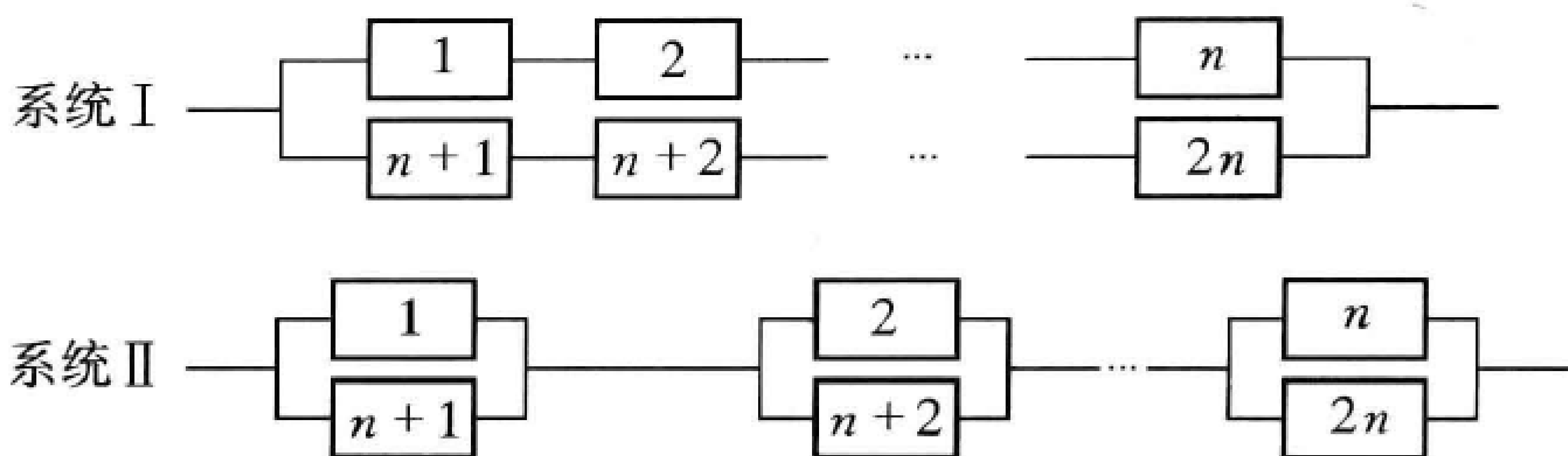
(1) 当 A 与 B 独立时,A 与 B 相容;

(2) 当 A 与 B 不相容时,A 与 B 独立。

7. 已知事件 A, B, C 相互独立,求证:$A\cup B$ 与 C 也独立。

8. 甲、乙、丙三台机床独立工作,在同一段时间内它们不需要工人照顾的概率分别为 0.7,0.8 和 0.9,求在这段时间内,最多只有一台机床需要人照顾的概率。

9. 如果构成系统的每个元件能正常工作的概率为 p(称为元件的可靠性),假设元件能否正常工作是相互独立的,分析下面各系统的可靠性。



10. 10 张奖券中含有 3 张中奖的奖券，每人购买 1 张，求

(1) 前三人中恰有一人中奖的概率；

(2) 第二人中奖的概率。

11. 在肝癌诊断中，有一种甲胎蛋白法，用这种方法能够检查出的肝癌患者中 95% 为真实患者，但也有可能将 10% 的正常人误诊。根据以往的记录，每 10000 人中有 4 人患有肝癌，试求：

(1) 某人经此检验法诊断患有肝癌的概率；

(2) 已知某人经此检验法检验患有肝癌，而他确实是肝癌患者的概率。

12. 一大批产品的优质品率为 30%，每次任取 1 件，连续抽取 5 次，计算下列事件的概率：

(1) 取到的 5 件产品中恰有 2 件是优质品；

(2) 在取到的 5 件产品中已发现有 1 件是优质品，这 5 件中恰有 2 件是优质品。

13. 每箱产品有 10 件，其次品数从 0 到 2 是等可能的。开箱检验时，从中任取 1 件，如果检验是次品，则认为该箱产品不合格而拒收。假设由于检验有误，1 件正品被认为是次品的概率是 2%，1 件次品被认为是正品的概率是 5%，试计算：

(1)抽取的 1 件产品为正品的概率；

(2)该箱产品通过验收的概率。

14. 假设一厂家生产的仪器，以概率 0.70 可以直接出厂，以概率 0.30 需进一步调试，经调试后以概率 0.80 可以出厂，并以概率 0.20 定为不合格品不能出厂。现该厂新生产了 $n(n\geqslant 2)$ 台仪器(假设各台仪器的生产过程相互独立)，求：

(1)全部能出厂的概率；

(2)其中恰有 2 件不能出厂的概率；

(3)其中至少有 2 件不能出厂的概率。

15. 进行一系列独立试验，每次试验成功的概率均为 p，试求以下事件的概率：

(1)直到第 r 次才成功;

(2)第 r 次成功之前恰失败 k 次;

(3) 在 n 次中取得 $r(1\leqslant r\leqslant n)$ 次成功;

(4) 直到第 n 次才取得 $r(1\leqslant r\leqslant n)$ 次成功。

16. 对飞机进行 3 次独立射击,第一次射击命中率为 0.4,第二次为 0.5,第三次为 0.7。击中飞机一次而飞机被击落的概率为 0.2,击中飞机两次而飞机被击落的概率为 0.6,若被击中三次,则飞机必被击落。求射击三次飞机被击落的概率。

第2章　随机变量及其分布

前面我们只是用初等数学的方法对试验中单个的事件进行了研究。而试验中的事件是很多的，如何从整体上把握这个试验，只有前面的知识是远远不够的。为此，我们引入随机变量。

2.1　随机变量

我们知道，许多试验的结果都是用数值表示。例如，掷一颗骰子出现的点数为:1,2,3,4,5,6;一箱中有10个产品，其中3个次品，7个正品，一次拿5个产品，其中的次品数为:0,1,2,3等等。它们有一个共同的规律，即不同的结果(样本点)对应不同的数值，这样我们就自然得到一个以样本点(ω)为自变量，以样本空间(Ω)为定义域的实函数。

定义2.1.1　设E为随机试验，它的样本空间为$\Omega=\{\omega\}$，如果对于每一个$\omega\in\Omega$均有实数$X(\omega)$与之对应，则称这个定义在Ω上的实单值函数$X(\omega)$为**随机变量(random variable)**，简记$X(\omega)$为X。

当然，有些试验的结果不具有数量的含意。例如，掷一枚硬币观察出现正、反面的情况。但我们可以引入变量按下列规定取值

$$X=\begin{cases}1, & \text{当出现正面时}\\ 0, & \text{当出现反面时}\end{cases}$$

这样X是一个随机变量，它取何值是由试验结果而定的。

在上面的掷骰子、抽检产品、掷硬币等试验中，对于随机变量的取值，我们可以按一定的顺序一一列出，这样的随机变量称为**离散型随机变量(discrete random variable)**。

有的随机变量，它可以取某一区间或某几个区间内的一切值，不可能一一列出。例如:某公共汽车每隔15分钟来一辆，那么某乘客在车站的候车时间X是一个随机变量，它可以取区间[0,15]内的一切值。某一自动装置的使用寿命X是一个随机变量，它可以取区间$(0,+\infty)$内的一切值。像这种可以取区间内一切值的随机变量称为**连续型随机变量(continuous random variable)**。

此外,若 X 是一个随机变量,那么以 X 为自变量构成的函数 $Y=f(X)$ 也是随机变量,称 Y 为随机变量 X 的函数。例如,由于各种因素的影响,一个球体直径的测量值 D 是一个随机变量,那么球体的体积 $V=\frac{4}{3}\pi\left(\frac{D}{2}\right)^3$ 是 D 的函数,显然 V 也是随机变量。

随机变量是随机事件的必然延伸与升华,随机事件是随机变量取特定值时的具体体现。

2.2 离散型随机变量及其分布列

掷一颗均匀的骰子,我们设 X 代表出现的点数,则 X 可能取到的值为:1,2,3,4,5,6。由古典概率可知,X 取各值的概率都等于$\frac{1}{6}$,列表如下:

X	1	2	3	4	5	6
P	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

这个表从概率的角度指出了随机变量在随机试验中取值的分布状况,称为 X 的概率分布。

一般地,设离散型随机变量 X 可能取的值为:x_1, x_2, …, x_i, …。X 取每一个值 $x_i(i=1, 2, \cdots)$的概率 $P(X=x_i)=p_i$,则称下表

X	x_1	x_2	…	x_i	…
P	p_1	p_2	…	p_i	…

为随机变量 X 的概率分布(probability disribution),也称为 **X 的分布列(distribution series)**。

显然,概率分布必须满足下列两个条件:

(1)$0\leqslant p_i\leqslant 1 \quad i=1, 2, \cdots$;

(2) $\sum_{i=1}^{\infty} p_i = 1$ 。

概率分布的重要作用是:对任意的 $a<b$,有

$$P(a < X \leqslant b) = \sum_{a < x_i \leqslant b} p_i$$

这样，我们一旦有了随机变量 X 的概率分布，就可以求得该随机变量 X 所代表的随机试验中任何事件的概率，从而对整个试验了如指掌。

例如在掷骰子的试验中

$$P(2 < X \leqslant 5) = P(X = 3) + P(X = 4) + P(X = 5) = \frac{1}{2},$$

$$P(-2 \leqslant X < 3.5) = P(X = 1) + P(X = 2) = \frac{1}{3}$$

不同的随机试验对应着不同的概率分布，下面我们介绍三种常见的离散型随机变量的概率分布。

1. 两点分布

定义 2.2.1　如果随机变量 X 的概率分布为

$$P(X = 1) = p;\quad P(X = 0) = 1 - p = q \qquad (0 < p < 1)$$

则称 **X 服从两点分布(two-point distribution)**，也称 0－1 **分布(0－1 distribution)**。

两点分布指的是一般只有两个结果的试验。例如掷硬币只有正面和反面，检查产品只有正品和次品，出生小孩只有男孩和女孩，神经细胞只有抑制和兴奋，系统电路只有通和不通……。

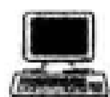
2. 二项分布

定义 2.2.2　如果随机变量 X 的概率分布为

$$P(X = k) = C_n^k p^k q^{n-k} \qquad k = 0, 1, 2, \cdots, n$$

其中 $0<p<1, q=1-p$，则称 X 服从参数为 n, p 的**二项分布(two-term distrbution)**，记为 $X \sim B(n, p)$。

n 重贝努里试验正是二项分布的试验背景。一个篮球运动员在 100 次投篮中投中的次数，n 台相互独立的同型号的设备同时开动时出现故障的设备数，产品检验中抽得的次品数等等，都是二项分布模型的具体化。两点分布是二项分布的特殊情况。

 **二项分布演示实验**

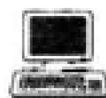
3. 泊松分布

定义 2.2.3 如果随机变量 X 的概率分布为

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda} \qquad k=0,1,2,\cdots,\lambda>0$$

则称 X 服从参数为 λ 的**泊松分布(Poisson distrbution)**,记为 $X\sim P(\lambda)$。

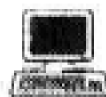
在实际中,有很多"排队"问题都可以近似地用泊松分布来描绘。如某段时间内候车室内旅客人数;电话交换台接到的呼叫次数;放射性分裂落到一个区域内的质点数目等等,都近似地服从泊松分布。

 **泊松分布演示实验**

容易验证,以上三种分布均满足概率分布的两个必要条件(读者可自行验证)。

可以证明(称为泊松定理),当 n 很大,p 很小,$\lambda=np$ 是一个不太大的常数时,可以用泊松分布做为二项分布的近似,即

$$C_n^k p^k (1-p)^{n-k} \approx \frac{\lambda^k}{k!}e^{-\lambda} \qquad k=0,1,2,\cdots,n$$

 **泊松定理演示实验**

例 2.2.1 某射手有 5 发子弹,每射一次命中的概率为 0.9,如果命中了就停止射击,否则一直射到子弹用尽为止,求耗用子弹数 X 的概率分布。

解 X 的所有可能取值为:1,2,3,4,5。

$P(X=1)=0.9$,

$P(X=2)=0.1\times 0.9=0.09$,

$P(X=3)=0.1^2\times 0.9=0.009$,

$P(X=4)=0.1^3\times 0.9=0.0009$,

$P(X=5)=0.1^4\times 0.9+0.1^5=0.0001$。

列表表示为:

X	1	2	3	4	5
P	0.9	0.09	0.009	0.0009	0.0001

当然也可以这样计算 $P(X=5)$，即

$$P(X=5)=1-P(X=1)-P(X=2)-P(X=3)-P(X=4)$$

例 2.2.2　某车间有 5 台同型号的机床，每台机床配备的电动机功率为 10 千瓦。已知每台机床工作时平均每小时实际开动 20 分钟，且开动与否是相互独立的。现因电力供应紧张，供电部门仅提供 30 千瓦的电力给这 5 台机床，问这 5 台机床能正常工作的概率有多大？

解　设 A=“机床在实际开动”，X=“同时实际开动的机床总数”，则 $P(A)=\frac{1}{3}$，且 X 服从 $n=5,p=\frac{1}{3}$ 的二项分布。由于每台电动机的功率为 10 千瓦，所以同时开动的车床数不能超过 $\frac{30}{10}=3$ 台。故所求概率为

$$\begin{aligned}P(X\leqslant 3)&=P(X=0)+P(X=1)+P(X=2)+P(X=3)\\&=C_5^0\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^5+C_5^1\left(\frac{1}{3}\right)^1\left(\frac{2}{3}\right)^4+C_5^3\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^3\\&\quad+C_5^3\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^2=\frac{232}{243}\\&\approx 0.95\end{aligned}$$

下面我们通过一个例子说明怎样求离散型随机变量函数的分布列。

例 2.2.3　设 X 的概率分布为

X	-1	0	1	2
P	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{10}$	$\frac{3}{10}$

求(1)$Y=X-1$ 的概率分布；(2)$Y=X^2$ 的概率分布。

解　(1)由 $Y=X-1$ 可知 Y 的所有可能取值为：$-2,-1,0,1$，且

$P(Y=-2)=P(X=-1)=\frac{1}{5}$；　　$P(Y=-1)=P(X=0)=\frac{2}{5}$；

$P(Y=0)=P(X=1)=\frac{1}{10}$；　　$P(Y=1)=P(X=2)=\frac{3}{10}$

故 $Y=X-1$ 的概率分布为

Y	-2	-1	0	1
P	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{10}$	$\frac{3}{10}$

(2)同理 $Y=X^2$ 的概率分布为

Y	0	1	4
P	$\frac{2}{5}$	$\frac{3}{10}$	$\frac{3}{10}$

其中 $P(Y=1)=P(X=-1)+P(X=1)=\frac{1}{5}+\frac{1}{10}=\frac{3}{10}$。

练习 2.2

1. 设 X 为随机变量,且 $P(X=k)=\frac{1}{2^k}$, $k=1, 2, \cdots$,则

(1) 判断上面的式子是否为 X 的概率分布;

(2) 若是,试求 $P(X$ 为偶数$)$和 $P(X\geqslant 5)$。

2. 已知 $P(X=k)=\frac{C\lambda^k}{k!}e^{-\lambda}, k=1, 2, \cdots$,且 $\lambda>0$,求 C。

3. 设一次试验成功的概率为 $p(0<p<1)$,不断进行重复试验,直到首次成功为止。用随机变量 X 表示试验的次数,求 X 的概率分布。

4. 设自动生产线在调整以后出现废品的概率为 p,当生产过程中出现废品时立即进行调整,代表在两次调整之间生产的合格品数,试求

(1)X 的概率分布;　　(2)$P(X\geqslant 5)$。

5. 一张考卷上有 5 道选择题,每道题列出 4 个可能答案,其中有一个答案是正确的。求某学生靠猜测能答对至少 4 道题的概率是多少?

6. 为了保证设备正常工作,需要配备适当数量的维修人员。根据经验每台设备发生故障的概率为 0.01,各台设备工作情况相互独立。

(1)若由 1 人负责维修 20 台设备,求设备发生故障后不能及时维修的概率;

(2)设有设备 100 台,1 台发生故障由 1 人处理,问至少需配备多少维修人员,才能保证设备发生故障而不能及时维修的概率不超过 0.01?

7. 设随机变量 X 服从参数为 λ 的 Poisson 分布,且 $P(X=0)=\frac{1}{2}$,求

(1)λ;　　(2)$P(X>1)$。

8. 设书籍上每页的印刷错误服从 Poisson 分布。经统计发现在某本书上,有一个印刷错误与有两个印刷错误的页数相同,求任意检验 4 页,每页上都没有印刷错误的概率。

9. 在长度为 t 的时间间隔内，某急救中心收到紧急呼救的次数 X 服从参数为 $\lambda=\frac{1}{2}t$ 的 Poisson 分布，而与时间间隔的起点无关(时间以小时计)，求

(1)某一天从中午 12 时至下午 3 时没有收到紧急呼救的概率；

(2)某一天从中午 12 时至下午 5 时至少收到 1 次紧急呼救的概率。

10. 已知 X 的概率分布为：

X	-2	-1	0	1	2	3
P	$2a$	$\frac{1}{10}$	$3a$	a	a	$2a$

试求(1) a；(2)$Y=X^2-1$ 的概率分布。

2.3　连续型随机变量及其密度

由于连续型随机变量的取值充满一个或若干个有限或无限区间，它的取值不可能一一列出，再用前面概率分布表的方法去描述连续型随机变量显然是不可能的。下面我们通过一个例子找出解决这个问题的最基本最原始的方法。

某厂生产某产品的规定尺寸为 25.40 mm，已知这批产品的最小尺寸为 25.20 mm，最大尺寸为 20.60 mm。现从这批产品中任取一产品，该产品的尺寸可看作连续型随机变量 X，它的取值范围为区间[25.20,25.60]。现从这批产品中任取 100 件，经测量就得到 100 个测量值。运用中学中有关的统计知识，可得到这 100 个数据的频率分布表(如表 2.3.1)和频率分布直方图(如图 2.3.1)。

表 2.3.1　频率分布表

分 组	频 数	频 率	累计频率
25.235～25.265	1	0.01	0.01
25.265～25.295	2	0.02	0.03
25.295～25.325	5	0.05	0.08
25.325～25.355	12	0.12	0.20
25.355～25.385	18	0.18	0.38
25.385～25.415	25	0.25	0.63

续表 2.3.1

分组	频数	频率	累计频率
25.415～25.445	16	0.16	0.79
25.445～25.475	13	0.13	0.92
25.475～25.505	4	0.04	0.96
25.505～25.535	2	0.02	0.98
25.535～25.565	2	0.02	1.00
合计	100	1.00	

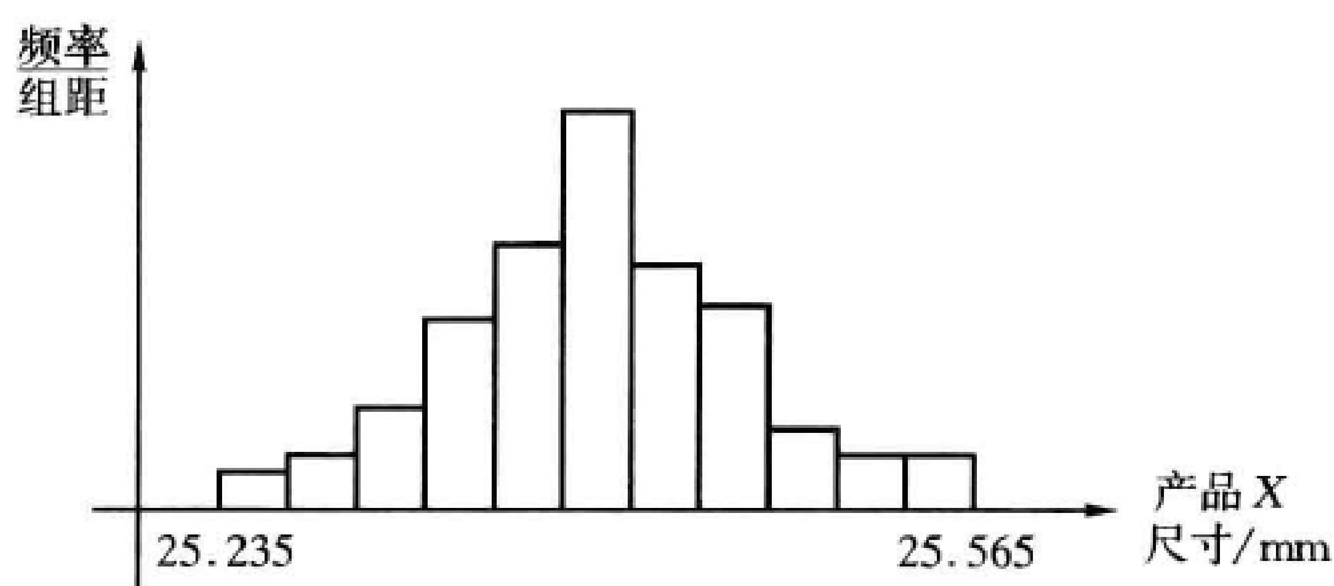


图 2.3.1

我们知道,频率是概率的近似值,每个小矩形的高低反映了随机变量 X 取值在各个小区间内的概率大小。随着所抽取的产品个数 n 的增大和所分数组的增加,图中所表示的频率就越接近于 X 在各小区间(越来越小)内取值的概率。由极限的知识可知,当 n 无限增大,分组的组距无限缩小,频率分布的直方图就会无限接近于一条光滑的曲线。这条曲线精确地反映了产品尺寸 X 在各个范围内取值的规律性(如图 2.3.2)。

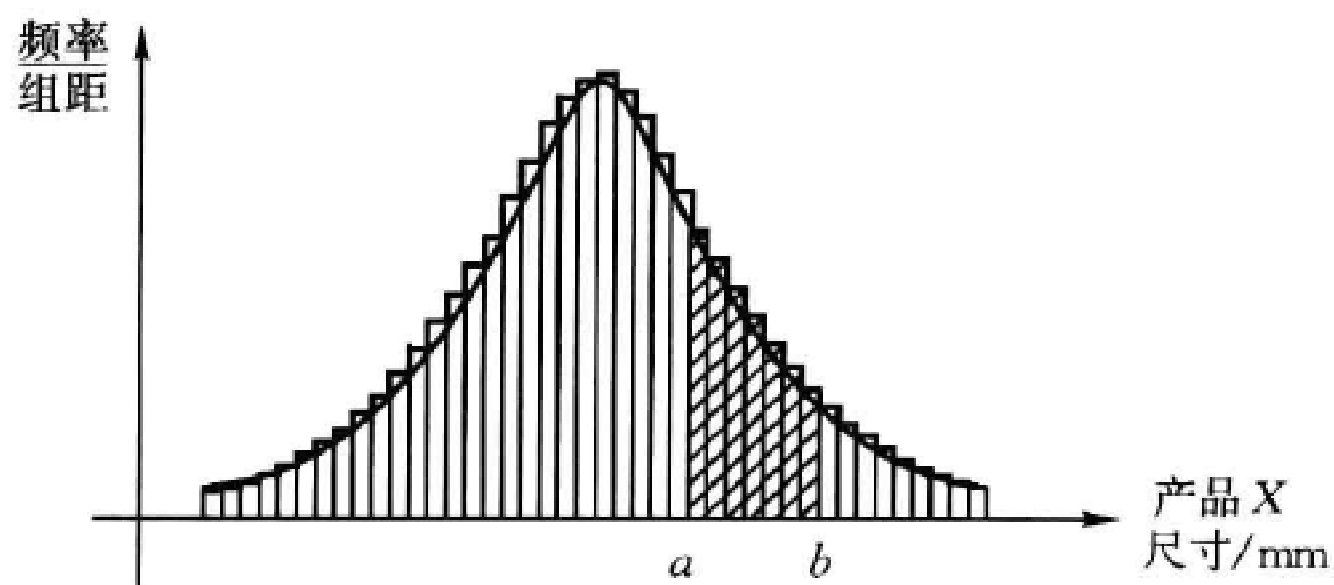


图 2.3.2

一般地,我们将这条曲线称为随机变量 X 的**概率密度曲线(curve of probabili-**

ty density)。以这条曲线为图形的函数称为随机变量 X 的**概率密度函数**(**function of probability density**),简称密度,记作 $f(x)$。

以上的过程类似于微积分中求曲边梯形面积的过程。它的基本思想都是先将连续问题离散化,然后通过无限细分取极限,将近似解转化为精确解。

显然,与离散型随机变量的概率分布类似,概率密度函数 $f(x)$必须具备以下两个条件:

(1) $f(x)\geqslant 0,\ -\infty<x<+\infty$;

(2) $\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$。

概率密度函数 $f(x)$的重要作用是:对任意的 $a<b$,有

$$P(a < X \leqslant b) = \int_a^b f(x)\mathrm{d}x$$

即图 2.3.2 中曲线下方阴影部分的面积。

由于连续型随机变量取值的连续性,它取值的有效集合必然是一个区间,即它在个别点上取值的概率为 0,因此密度函数 $f(x)$反映的是随机变量 X 取 x 邻域内值的概率的大小。用密度函数来描述连续型随机变量与用分布列来描述离散型随机变量是颇为相似的。这样,概率密度函数 $f(x)$就成为了解连续型随机变量的"窗口"和有力的工具。我们一旦有了 $f(x)$,就可以全面地掌握其相应的随机试验。

下面我们介绍三种常见的连续型随机变量的密度。

1. 均匀分布

定义 2.3.1　如果随机变量 X 的概率密度函数为

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \leqslant x \leqslant b \\ 0, & \text{其它} \end{cases}$$

则称 X 服从区间$[a,b]$上的**均匀分布**(**evenly distribution**),记作 $U[a,b]$。其概率密度曲线 $f(x)$如图 2.3.3 所示。

对于任意区间$[c,\ d]\subset[a,b]$,都有

$$P(c \leqslant X \leqslant d) = \int_c^d \frac{1}{b-a}\mathrm{d}x = \frac{d-c}{b-a}$$

由此可以看出,它在$[a,\ b]$上任何一个子区间取值的概率,与该子区间的长度成正比,与子区间在$[a,\ b]$中的位置无关。X 取值的等可能性反映了一种"均匀性",可以说这是一种"连续情形下的古典概型",在实际中有广泛的应用。前面谈到的乘客在车站的"候车时间 X"就服从均匀分布。随机取一个实数,只考虑其小

数部分，则小数部分也服从[0, 1]上的均匀分布。

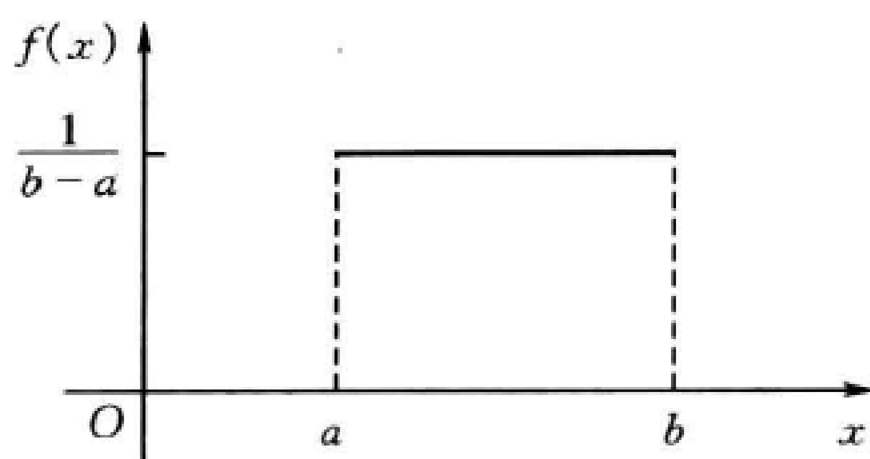


图 2.3.3

2. 指数分布

定义 2.3.2 如果随机变量 X 的概率密度函数为

$$f(x)=\begin{cases}\lambda e^{-\lambda x}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

其中 $\lambda>0$，则称 **X 服从参数为 λ 的指数分布(exponential distribution)**，记作 $E(\lambda)$。其概率密度曲线 $f(x)$ 如图 2.3.4 所示。

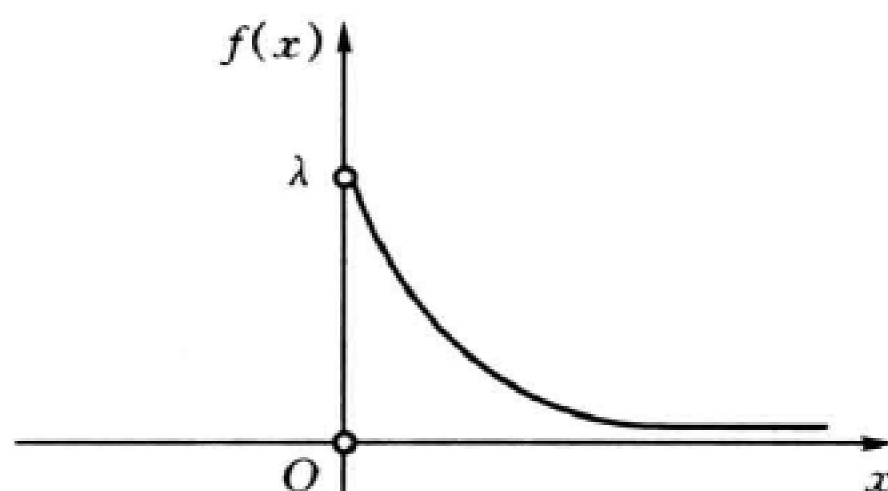


图 2.3.4

指数分布常用作各种"寿命"分布的近似。例如，无线电元件的寿命，动物的寿命，电话的通话时间等都近似服从指数分布。指数分布具有无记忆性，它的直观意义是有些元件在使用过程中损坏与否同过去使用的历史无关。

3. 正态分布

定义 2.3.3 如果随机变量 X 的概率密度函数为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ -\infty<x<+\infty$$

其中 $\sigma>0$，则称 **X 服从参数为 μ, σ^2 的正态分布(normal distribution)**。简记为 $N(\mu, \sigma^2)$。其概率密度曲线 $f(x)$ 如图 2.3.5 所示。

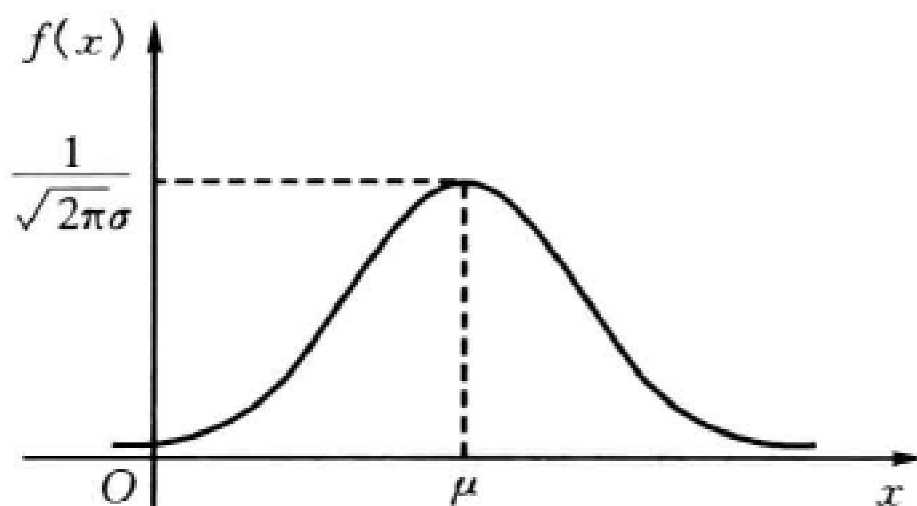


图 2.3.5

特别当 $\mu=0$, $\sigma=1$ 时，称随机变量 X 服从**标准正态分布(standard normal distribution)**，记为 $N(0, 1)$。此时 X 的概率密度函数记为 $\varphi(x)$，即

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}},\ -\infty<x<+\infty$$

密度函数曲线 $\varphi(x)$ 如图 2.3.6 所示。

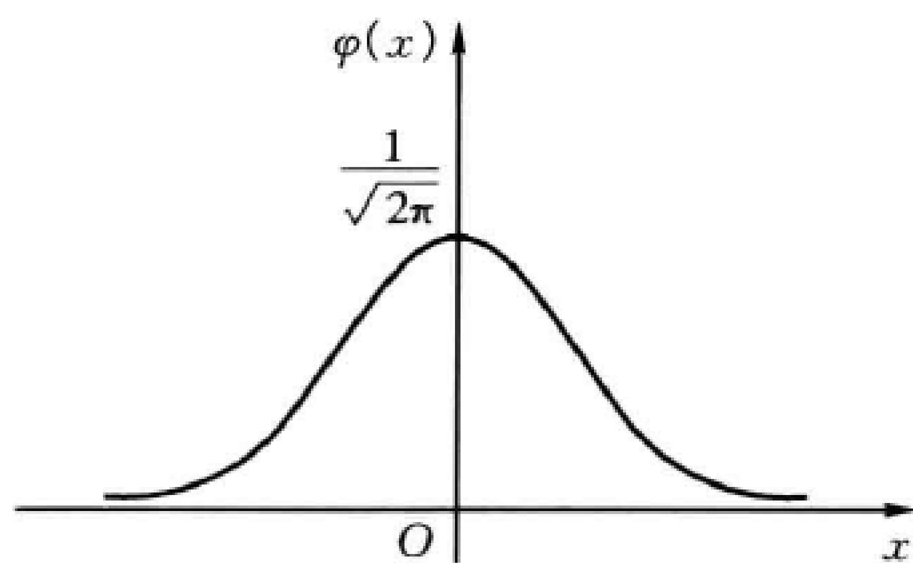


图 2.3.6

由微积分中的作图法及 $f(x)$ 的图形，可得出曲线 $f(x)$ 具有以下性质：

(1)以直线 $x=\mu$ 为对称轴；

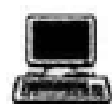
(2)以直线 $y=0$（x 轴）为渐近线；

(3)当 $x=\mu$ 时，$f(x)$ 有极大值 $\frac{1}{\sqrt{2\pi}\sigma}$；

(4) $\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$，即曲线与 x 轴之间的面积为 1。

以上 4 条性质中，当 $\mu=0$，$\sigma=1$ 时即为 $\varphi(x)$ 的性质。

正态分布在实践中和理论上都非常重要。测量误差和很多产品的物理指标（产品的长度、宽度、质量指标等等）都可以看作服从正态分布。在前面引进概率密度函数概念时遇到的产品尺寸就服从正态分布。另外其它许多分布在一定条件下都可看成正态分布。

 **一维正态曲线演示实验**

例 2.3.1　设连续型随机变量 X 的密度函数为

$$f(x) = \begin{cases} ax + b, & 0 < x < 2 \\ 0, & \text{其它} \end{cases}$$

且 $P(1<X<3)=0.25$，求常数 a 和 b；并计算 $P(X>1.5)$。

解　由密度函数的性质 $\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = \int_0^2 (ax+b)\mathrm{d}x = 2a+2b = 1$ 及 $P(1<X<3) = \int_1^3 f(x)\mathrm{d}x = \int_1^2 (ax+b)\mathrm{d}x = 1.5a+b = 0.25$，可得方程组

$$\begin{cases} 2a+2b=1 \\ 1.5a+b=0.25 \end{cases}$$

解此方程组得 $a=-0.5$，$b=1$。于是

$$\begin{aligned} P(X>1.5) &= \int_{1.5}^{+\infty} f(x)\mathrm{d}x \\ &= \int_{1.5}^{2} (-0.5x+1)\mathrm{d}x = 0.0625 \end{aligned}$$

例 2.3.2　设连续型随机变量 X 的密度函数为

$$f(x)=\begin{cases}\frac{1}{3}, & x\in[0,1]\\ \frac{2}{9}, & x\in[3,6]\\ 0, & \text{其它}\end{cases}$$

若 $P(X\geqslant k)=\frac{2}{3}$,求数值 k 的取值范围。

解 因为 $P(X\geqslant k)=\int_k^{+\infty}f(x)\mathrm{d}x$,所以

若 $k<0$,则

$$P(X\geqslant k)=\int_k^0 0\mathrm{d}x+\int_0^1\frac{1}{3}\mathrm{d}x+\int_3^6\frac{2}{9}\mathrm{d}x=\frac{1}{3}+\frac{2}{3}=1>\frac{2}{3};$$

若 $0\leqslant k<1$,则

$$P(X\geqslant k)=\int_k^1\frac{1}{3}\mathrm{d}x+\int_3^6\frac{2}{9}\mathrm{d}x=\frac{1}{3}(1-k)+\frac{2}{3}>\frac{2}{3};$$

若 $1\leqslant k\leqslant 3$,则

$$P(X\geqslant k)=\int_k^3 0\mathrm{d}x+\int_3^6\frac{2}{9}\mathrm{d}x=\frac{2}{3};$$

若 $3<k<6$,则

$$P(X\geqslant k)=\int_k^6\frac{2}{9}\mathrm{d}x=\frac{2}{9}(6-k)<\frac{2}{3};$$

若 $k\geqslant 6$,则

$$P(X\geqslant 6)=\int_k^{+\infty}0\mathrm{d}x=0;$$

故应取 $1\leqslant k\leqslant 3$。

此题由 $f(x)$的几何意义很容易求得(如图 2.3.7)。

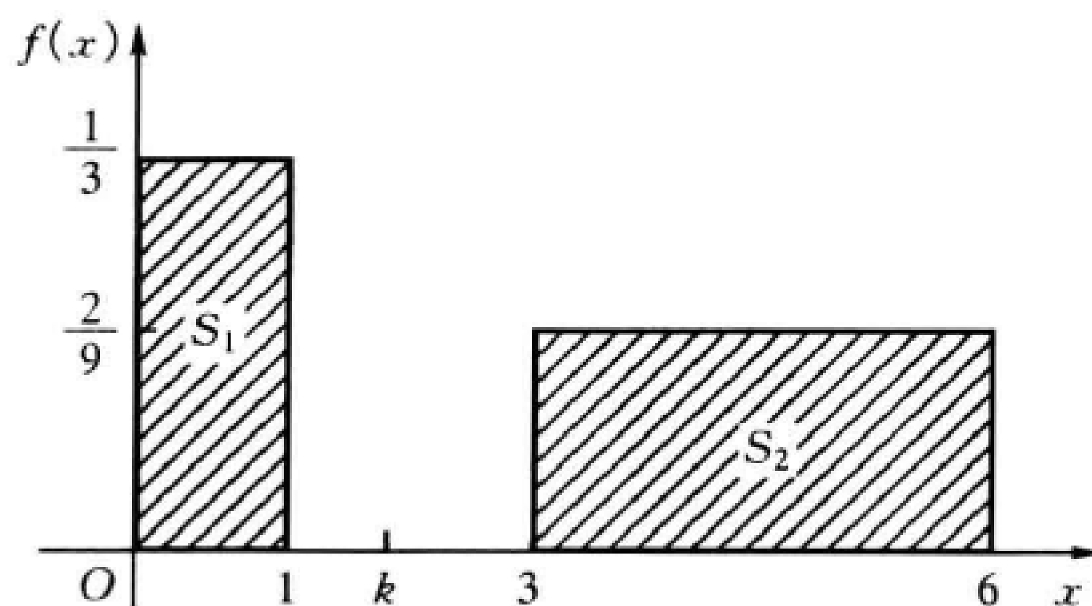


图 2.3.7

因为 $S_2=\frac{2}{9}\times 3=\frac{2}{3}$，而 $P(X\geqslant k)$ 表示的正是在点 $x=k$ 右侧的面积值，所以只有 $1\leqslant k\leqslant 3$ 时，才可使 $P(X\geqslant k)=\frac{2}{3}$。

练习 2.3

1. 设连续型随机变量 X 的概率密度曲线如图 2.3.8 所示，试求

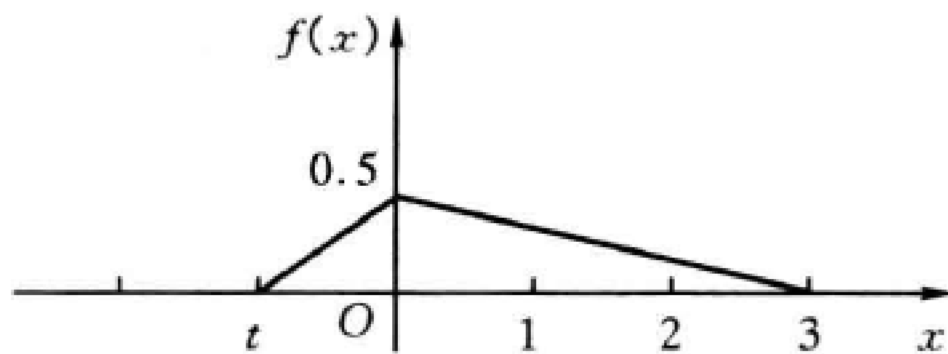


图 2.3.8

(1) t 的值；(2) X 的概率密度 $f(x)$；(3) $P(-2<X\leqslant 2)$。

2. 设连续型随机变量 X 的概率密度为

$$f(x)=\begin{cases}\sin x, & 0\leqslant x\leqslant a\\ 0, & \text{其它}\end{cases}$$

试确定常数 a，并求 $P(X>\frac{\pi}{6})$。

3. 函数 e^{-x^2+x} 乘以什么常数可使其变成概率密度？

4. 随机变量 $X\sim N(\mu,\sigma^2)$，其概率密度函数为

$$f(x)=\frac{1}{\sqrt{6\pi}}e^{-\frac{x^2-4x+4}{6}}\quad -\infty<x<+\infty$$

试求 μ,σ^2；若已知 $\int_C^{+\infty}f(x)\mathrm{d}x=\int_{-\infty}^{C}f(x)\mathrm{d}x$，求 C。

5. 设连续型随机变量 X 的概率密度为

$$f(x)=\begin{cases}2x, & 0\leqslant x\leqslant 1\\ 0, & \text{其它}\end{cases}$$

以 Y 表示对 X 的三次独立重复试验中"$X\leqslant\frac{1}{2}$"出现的次数，试求概率 $P(Y=2)$。

6. 设随机变量 X 服从 $[0,5]$ 上的均匀分布，试求 $P(x_1<X<x_2)$，如果

(1) $x_1<1<x_2<5$；(2) $1<x_1<5<x_2$

7. 设顾客排队等待服务的时间 X（以分计）服从 $\lambda=\frac{1}{5}$ 的指数分布。某顾客等

待服务，若超过 10 分钟，他就离开。他一个月要去等待服务 5 次，以 Y 表示一个月内他未等到服务而离开的次数，试求 Y 的概率分布和 $P(Y\geqslant 1)$。

2.4 分布函数

1. 分布函数

离散型随机变量可以由概率分布来描述，连续型随机变量可以由密度函数来描述。其它类型的随机变量如何来描述呢？分布函数就是一种描述包括离散型和连续型随机变量在内的一切类型随机变量的非常有效的工具。

定义 2.4.1 设 X 是一个随机变量，对于任意实数 x，函数

$$F(x)=P(X\leqslant x),\ -\infty<x<+\infty$$

称为随机变量 X 的**分布函数(distribution function)**。

对于任意给定的 x，$(X\leqslant x)$是一个随机事件，而 $F(x)$就是这一事件发生的概率。因此对任意的 $a<b$，有

$$P(a<X\leqslant b)=P(X\leqslant b)-P(X\leqslant a)=F(b)-F(a)。$$

从这个意义上讲，分布函数 $F(x)$完整地描述了随机变量(无论是离散的还是连续的)的统计规律。

分布函数 $F(x)$具有以下性质：

(1) $F(x)$是单调不减的，即对于任意 $x_1<x_2$，有 $F(x_1)\leqslant F(x_2)$；

(2) $0\leqslant F(x)\leqslant 1$；

(3) $\lim\limits_{x\to-\infty}F(x)=0,\ \lim\limits_{x\to+\infty}F(x)=1$，记作 $F(-\infty)=0,\ F(+\infty)=1$；

(4) $F(x)$在每点 x 处均是右连续的，即 $\lim\limits_{\varepsilon\to 0^+}F(x+\varepsilon)=F(x)$，记作 $F(x+0)=F(x)$。

对于离散型随机变量 X，相应的分布函数性质如下。

(1) 若已知 X 的概率分布为 $P(X=x_k)=p_k, k=1, 2,\cdots$，那么随机变量 X 的分布函数 $F(x)=\sum\limits_{x_k<x}p_k$；反之，若已知 X 的分布函数为 $F(x)$，则随机变量 X 的概率分布为 $p_k=P(X=x_k)=F(x_k)-F(x_k-0),\ k=1, 2, \cdots$。

所以概率分布和分布函数是一一对应的。

(2) 对任意的 $a<b$，有

$P(X<a)=F(a-0),\ P(X\geqslant a)=1-F(a-0)$。

$P(a\leqslant X\leqslant b)=F(b)-F(a)$；

$P(a\leqslant X<b)=F(b-0)-F(a-0)$；$P(a<X<b)=F(b-0)-F(a)$。

对于连续型随机变量 X,相应的分布函数性质表现如下。

(1) 若已知 X 的概率密度函数为 $f(x)$,那么随机变量 X 的分布函数

$$F(x)=\int_{-\infty}^{x} f(t)\mathrm{d}t,\ -\infty<x<+\infty$$

反之,若已知 X 的分布函数为 $F(x)$,在密度函数连续的点 x 处,有 $F'(x)=f(x)$,$F(x)$为 x 的连续函数。

所以概率密度函数和分布函数是一一对应的。

(2) 对任意的 $a<b$,有

$$P(X<a)=P(X\leqslant a)=F(a);$$
$$P(X\geqslant a)=P(X>a)=1-F(a)$$
$$P(a\leqslant X\leqslant b)=P(a\leqslant X<b)=P(a<X\leqslant b)$$
$$=P(a<X<b)=F(b)-F(a);$$

例 2.4.1　设 X 的概率分布为:

X	-1	0	2
P	0.3	0.5	0.2

求①分布函数 $F(x)$在 $x=1.5$ 处的值;②X 的分布函数并作图;③$P(X>0)$和 $P(0.5<X<5)$。

解　① $F(1.5)=P(X\leqslant 1.5)=P(X=-1)+P(X=0)=0.3+0.5=0.8$

② 当 $x\in(-\infty,\ -1)$时,$F(x)=0$;

当 $x\in[-1,\ 0)$时,$F(x)=P(X=-1)=0.3$;

当 $x\in[0,\ 2)$时,$F(x)=P(X=-1)+P(X-0)=0.8$;

当 $x\in[2,\ +\infty]$时,$F(x)=P(X=-1)+P(X=0)+P(X=2)=1$。

综上所知,X 的分布函数 $F(x)$为

$$F(x)=\begin{cases}0, & x<-1\\ 0.3, & -1\leqslant x<0\\ 0.8, & 0\leqslant x<2\\ 1, & x\geqslant 2\end{cases}$$

$F(x)$的图形如图 2.4.1 所示。

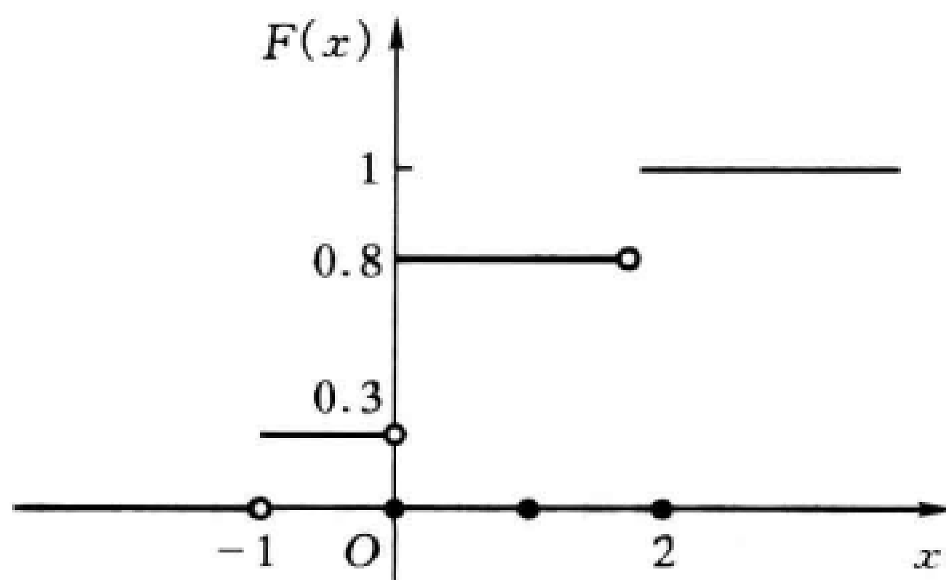


图 2.4.1

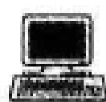
由 $F(x)$的图形,可以很直观地显示出 $F(x)$的 4 条基本性质。

③ $P(X>0)=1-F(0)=1-0.8=0.2$,当然也可以这样计算 $P(X>0)=$

$P(X=2)=0.2$。

$P(0.5<X<5)=F(5-0)-F(0.5)=1-0.3=0.7$，当然也可以这样计算 $P(0.5<X<5)=P(X=0)+P(X=2)=0.7$。

反过来，若已知 X 的分布函数 $F(x)$，如何求 X 的概率分布(请读者自己计算)。

 **离散型随机变量常见分布演示实验**

例 2.4.2 设随机变量 X 服从$[a, b]$上的均匀分布，求①X 的分布函数$F(x)$并作图；②若 $x_1<a<x_2<b$，求 $P(x_1<X<x_2)$。

解 因为 X 的密度函数为

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & a\leqslant x\leqslant b\\ 0, & \text{其它}\end{cases}$$

当 $x\in(-\infty, a)$ 时，$F(x)=\int_{-\infty}^{x}f(t)\mathrm{d}t=\int_{-\infty}^{x}\mathrm{d}t=0$；

当 $x\in[a, b)$ 时，$F(x)=\int_{-\infty}^{x}f(t)\mathrm{d}t=\int_{a}^{x}\dfrac{1}{b-a}\mathrm{d}t=\dfrac{x-a}{b-a}$；

当 $x\in[b, +\infty]$ 时，$F(x)=\int_{-\infty}^{x}f(t)\mathrm{d}t=\int_{a}^{b}\dfrac{1}{b-a}\mathrm{d}t=1$；

综上所知，X 的分布函数 $F(x)$ 为

$$F(x)=\begin{cases}0, & x<a\\ \dfrac{x-a}{b-a}, & a\leqslant x<b\\ 1, & x\geqslant b\end{cases}$$

$F(x)$的图形如图 2.4.2 所示。

由 $F(x)$的图形，可以验证 $F(x)$的 4 条基本性质。

② $P(x_1<X<x_2)=F(x_2)-F(x_1)=\dfrac{x_2-a}{b-a}-0=\dfrac{x_2-a}{b-a}$。

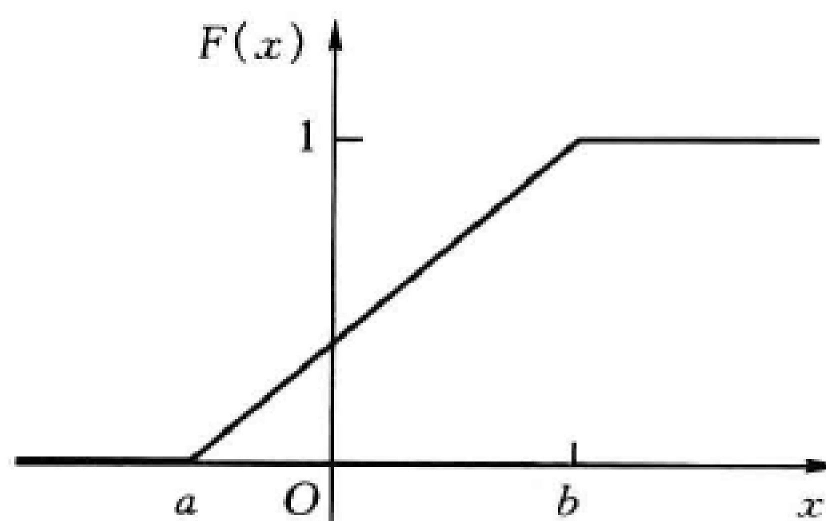
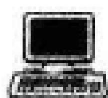


图 2.4.2

 **连续型随机变量常见分布演示实验**

例 2.4.3 已知连续型随机变量 X 的分布函数为

$$F(x)=A+B\arctan x,\ -\infty<x<+\infty$$

试求① A 和 B；②概率密度函数 $f(x)$。

解　①因为 $F(-\infty)=A-B\cdot\frac{\pi}{2}=0$，$F(+\infty)=A+B\cdot\frac{\pi}{2}=1$，解方程组可得 $A=\frac{1}{2}$，$B=\frac{1}{\pi}$。

② $f(x)=F'(x)=\dfrac{1}{\pi(1+x^2)}$，$-\infty<x<+\infty$

2. 正态分布的分布函数

在日常生活中，无论是实际计算还是理论分析，正态分布起着重要的作用，为此作为讨论分布函数的例子，我们对正态分布的分布函数作进一步的研究。

设 $X\sim N(\mu,\sigma^2)$，其概率密度函数为 $f(x)$，则 X 的分布函数为

$$\begin{aligned}F(x)&=\int_{-\infty}^{x}f(t)\,\mathrm{d}t\\&=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(t-\mu)^2}{2\sigma^2}}\mathrm{d}t,\ -\infty<x<+\infty\end{aligned}$$

若 $X\sim N(0,1)$。则标准正态分布的分布函数为

$$\begin{aligned}\Phi(x)&=\int_{-\infty}^{x}\varphi(t)\,\mathrm{d}t\\&=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t,\ -\infty<x<+\infty\end{aligned}$$

那么 $F(x)$ 和 $\Phi(x)$ 之间有什么关系呢？

因为 $F(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(t-\mu)^2}{2\sigma^2}}\mathrm{d}t$，由定积分的换元法，令 $y=\frac{t-\mu}{\sigma}$，于是 $\mathrm{d}t=\sigma\cdot\mathrm{d}y$，从而

$$F(x)=\int_{-\infty}^{\frac{x-\mu}{\sigma}}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{y^2}{2}}\mathrm{d}y=\Phi\left(\frac{x-\mu}{\sigma}\right)$$

所以便得到如下结论：若 $X\sim N(\mu,\sigma^2)$，则对任意的 $a<b$，有

$$\begin{aligned}P(a<X\leqslant b)&=F(b)-F(a)\\&=\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)\end{aligned}$$

由 $\Phi(x)$ 的对称性知：$\Phi(+\infty)=0$，$\Phi(0)=\frac{1}{2}$，$\Phi(-x)=1-\Phi(x)$。

例 2.4.4　已知连续型随机变量 $X\sim N(10,2^2)$，求 $P(10<X<13)$，$P(X\geqslant13)$，$P(|X-10|<2)$。

解 $P(10<X<13)=\Phi\left(\frac{13-10}{2}\right)-\Phi\left(\frac{10-10}{2}\right)=\Phi(1.5)-\Phi(0)$

$$=0.9332-0.5=0.4332;$$

$$P(X\geqslant 13)=1-\Phi\left(\frac{13-10}{2}\right)=1-0.9332=0.0668;$$

$$P(|X-10|<2)=P(8<X<12)=\Phi\left(\frac{12-10}{2}\right)-\Phi\left(\frac{8-10}{2}\right)$$

$$=2\Phi(1)-1=2\times 0.8413-1=0.6826$$

例 2.4.5 设连续型随机变量 $X\sim N(\mu,\sigma^2)$，且已知 $P(X\leqslant -1.6)=0.036$，$P(X>5.9)=0.242$，求 u，σ，$P(X>0)$。

解 因 $P(X\leqslant -1.6)=\Phi\left(\frac{-1.6-\mu}{\sigma}\right)=0.036$，故 $\Phi\left(\frac{1.6+\mu}{\sigma}\right)=1-0.036=$ 0.964。

又 $P(X>5.9)=1-P(X\leqslant 5.9)=1-\Phi\left(\frac{5.9-\mu}{\sigma}\right)=0.242$，所以 $\Phi\left(\frac{5.9-\mu}{\sigma}\right)=$ 0.758。查表可得

$$\begin{cases}\dfrac{1.6+\mu}{\sigma}=1.8\\ \dfrac{5.9-\mu}{\sigma}=0.7\end{cases}\qquad 即\begin{cases}\mu=3.8\\ \sigma=3\end{cases}$$

$$P(X>0)=1-\Phi\left(\frac{0-3.8}{3}\right)=1-\Phi(-1.27)=\Phi(1.27)=0.898$$

3. 连续型随机变量函数的分布

下面通过一个例子来说明怎样求连续型随机变量函数的概率密度函数。

例 2.4.6 设随机变量 $X\sim N(\mu,\sigma^2)$，求 $Y=kX+b$ 的概率密度函数，其中 k，b 为常数，$k\neq 0$。

解 此类型题目一般先求出 Y 的分布函数，然后通过求导数 $F'_Y(y)$，求出密度函数 $f_Y(y)$。

设 $k>0$，Y 的分布函数为 $F_Y(y)$，则

$$F_Y(y)=P(Y\leqslant y)=P(kX+b\leqslant y)=P\left(x\leqslant \frac{y-b}{k}\right)$$

$$=\int_{-\infty}^{\frac{y-b}{k}} f_X(x)\mathrm{d}x$$

故 $Y=kX+b$ 的概率密度函数为

$$f_Y(y)=F'_Y(y)=f_X\left(\frac{y-b}{k}\right)\cdot\frac{1}{k}$$

$$= \frac{1}{\sqrt{2\pi}(k\sigma)} \exp\left\{-\frac{[y-(b+k\mu)]^2}{2(k\sigma)^2}\right\} \quad (k>0);$$

同理可得，$k<0$ 时，

$$f_Y(y) = \frac{1}{\sqrt{2\pi}(-k\sigma)} \exp\left\{-\frac{[y-(b+k\mu)]^2}{2(k\sigma)^2}\right\}$$

综上所知，当 $k\neq 0$ 时，有

$$f_Y(y) = \frac{1}{\sqrt{2\pi}(|k|\sigma)} \exp\left\{-\frac{[y-(b+k\mu)]^2}{2(k\sigma)^2}\right\}$$

这表明 $Y=kX+b\sim N(b+k\mu,\ k^2\sigma^2(k\neq 0))$，也就是说，服从正态分布的随机变量的线性函数仍服从正态分布。此方法称为“分布函数法”，是求连续型随机变量函数分布的一般化方法，这里我们再加以概括总结如下：

首先，根据分布函数定义，从随机变量函数的分布函数出发，找出函数和自变量分布函数之间的关系，这一步的关键是解不等式；

其次，求导数，找出函数和自变量密度之间的关系，这一步的关键是复合函数求导数；

最后，将自变量的密度具体化。这一步的关键是知道内层函数、外层函数，求复合函数，有时需要分类讨论。

下面不加证明地给出当 $y=g(x)$ 为单调可导函数时的一般结论。

设 $X\sim f_X(x)$，$y=g(x)$ 是单调可导函数，其导数恒不为零。记 $x=h(y)$ 是 $y=g(x)$ 的反函数，$(a,\ b)$ 是 $y=g(x)$ 的值域，其中 $-\infty<a<x<b<+\infty$，则 $Y=g(X)$ 是连续型随机变量，其密度为

$$f_Y(y) = \begin{cases} f_X[h(y)]\cdot|h'(y)|, & a<y<b \\ 0, & \text{其它} \end{cases}$$

显然例 2.4.6 可直接用此结论来计算。

练习 2.4

1. 已知随机变量 X 的概率分布为 $P(X=1)=0.2$，$P(X=2)=0.3$，$P(X=3)=0.5$，试求 X 的分布函数；$P(0.5\leqslant X\leqslant 2)$；画出 $F(x)$ 的曲线。

2. 设连续型随机变量 X 的分布函数为

$$F(x) = \begin{cases} 0, & x<-1 \\ 0.4, & -1\leqslant x<1 \\ 0.8, & 1\leqslant x<3 \\ 1, & x\geqslant 3 \end{cases}$$

试求(1)X 的概率分布；(2)$P(X<2|X\neq1)$。

3. 从家到学校的途中有3个交通岗，假设在各个交通岗遇到红灯的概率是相互独立的，且概率均是0.4，设 X 为途中遇到红灯的次数，试求(1)X 的概率分布；(2) X 的分布函数。

4. 试求习题2.3中第1题 X 的分布函数，并画出 $F(x)$的曲线。

5. 设连续型随机变量 X 的分布函数为

$$F(x)=\begin{cases}A+Be^{-2x}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

试求(1)a, b 的值；(2)$P(-1<X<1)$；(3)概率密度函数 $f(x)$。

6. 设 X 为连续型随机变量，其分布函数如下

$$F(x)=\begin{cases}a, & x<1\\ bx\ln x+cx+d, & 1\leqslant x\leqslant \mathrm{e}\\ d, & x>\mathrm{e}\end{cases}$$

试确定 $F(x)$中的 a, b, c, d 的值。

7. 设随机变量 X 的概率密度函数为 $F(x)=\dfrac{a}{\pi(1+x^2)}$，试确定 a 的值并求 $F(x)$和 $P(|X|<1)$。

8. 假设某地在任何长为 t(年)的时间间隔内发生地震的次数 $N(t)$服从参数为 $\lambda=0.1t$ 的 Poisson 分布，X 表示连续两次地震之间相隔的时间(单位:年)。

(1)证明 X 服从指数分布并求出 X 的分布函数；

(2)今后3年内再次发生地震的概率；

(3)今后3年到5年内再次发生地震的概率。

9. 设 $X\sim N(-1,16)$，试计算(1)$P(X<2.44)$；(2)$P(X>-1.5)$；(3)$P(|X|<4)$；(4)$P(|X-1|>1)$。

10. 某科统考成绩 X 近似服从正态分布 $N(70,10^2)$，第100名的成绩为60分，问第20名的成绩约为多少分？

11. 设随机变量 X 和 Y 均服从正态分布，$X\sim N(\mu,4^2)$，$Y\sim N(\mu,5^2)$，而 $p_1=P(X\leqslant\mu-4)$，$p_2=P(Y\geqslant\mu+5)$，试证明：$p_1=p_2$。

12. 设随机变量 X 服从均匀分布，令 $Y=cX+d(c\neq0)$，试求随机变量 Y 的密度函数。

2.5 应用 SPSS 计算概率

SPSS 是美国 SPSS 公司自20世纪80年代初开发的大型统计分析软件包。

经过逐步升级，其功能更强大，应用更广泛，界面也十分友好，非常适合作为教学和科研的统计软件。本书使用 SPSS11.5 for Windows 版本进行有关统计分析和数据处理。

1. 正态分布

设 $X \sim N(0, 1)$，则 $P(X \leqslant x)$可以通过以下步骤计算。

在 SPSS 数据编辑窗口(需要随便打开一个数据文件)，依次点击 Tansform→Compute 得到一个对话框，在 Targat Variable 里写入 p，在右侧 Functions：中选 CDF. NORMAL (q, mean, stddev) 并点击向上的箭头使该函数放入 Numeric Expression：中，在三个问号处填入(x, 0, 1)，点击"OK"便得到概率 $P(X \leqslant x)$。如$P(X \leqslant 1.96)$= CDF. NORMAL(1.96,0,1)=0.975，$P(X \leqslant -1.96)$= CDF. NORMAL(−1.96,0,1)=0.025 ，$P(|X| \leqslant 1.96)$= CDF. NORMAL(1.96,0,1) − CDF. NORMAL(−1.96,0,1)=0.975−0.025=0.950。

对于一般正态分布 $N(\mu, \sigma^2)$，仅需要将 CDF. NORMAL(q, mean, stddev)中的 mean 写成 μ，stddev 写成 σ 即可。

注意，对于标准正态分布也可用函数 CDF. NORM (z value)，如 $X \sim N(0, 1)$，$P(X \leqslant 1.96)$= CDF. NORM (1.96)=0.975。

2. 二项分布

设 $X \sim B(n, p)$，则 $P(X \leqslant x)$=CDF. BINOM (x, n, p)。如 $X \sim B(5, 0.5)$，$P(X \leqslant 3)$=CDF. BINOM(3,5 ,0.5)=0.8125。

$P(X=3)=P(X \leqslant 3)-P(X \leqslant 2)$

=CDF. BINOM(3,5 ,0.5)− CDF. BINOM(2,5 ,0.5)

=0.8125−0.5000=0.3125。

或者 $P(X=3)$= PDF. BINOM(3,5,0.5)=0.3125。

3. 泊松分布

设 $X \sim P(\lambda)$，则 $P(X \leqslant x)$=CDF. POISSON (x, λ)。如 $X \sim P(0.9)$，

$P(X=2)=P(X \leqslant 2)-P(X \leqslant 1)$

= CDF. POISSON(2,0.9)−CDF. POISSON(1,0.9)

=0.9371−0.7725=0.1646。

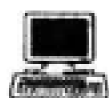
或者 $P(X=2)$=PDF. POISSON(2,0.9)=0.1646。

4. 指数分布

设 $X \sim E(\lambda)$，则 $P(X \leqslant x)=$CDF. EXP (x, λ)。

5. 均匀分布

设 $X \sim U[a, b]$，则 $P(X \leqslant x)=$ CDF. UNIFORM(x, a, b)。

 **常见分布函数值演示实验**

第 3 章　随机向量及其分布

第 2 章我们讨论了一维随机变量，但是在实际应用中，有一些随机现象用一个随机变量是无法进行全面描述的。例如为了评估某种电视机的质量，不仅要看"电视机在一年中发生的故障数"，而且还要同时考虑"一年中电视机使用的小时数"、"电视机画面的清晰度"等指标；冶炼钢水时，要同时考察其含碳量、含硫量以及某种稀有金属的含量等多个指标；在研究儿童的身体发育状况时，要同时分析身高、体重、视力等多个方面的因素；在研究国民经济发展的运行状况时，要经常分析居民的储蓄存款余额、国民收入及通货膨胀等多个经济指标。因此，同时研究多个随机变量的概率分布规律是十分必要的。它不仅可以使我们更全面、更准确地掌握所研究的随机现象，而且还可以对其中每一个随机变量的概率特征及这些变量间的联系有更清晰的认识。为此，我们需将这多个随机变量作为一个整体，看作一个**随机向量(random vector)**来分析，并研究其概率分布规律及这些变量间的相互依赖关系等。

3.1　随机向量及其分布

1. 随机向量

定义 3.1.1　以 n 个随机变量 $X_1, X_2, \cdots, X_n$ 为分量的 n 维向量 $(X_1, X_2, \cdots, X_n)$ 称为 **n 维随机向量**。

以下主要研究二维随机向量的情形。

2. 二维离散型随机向量及其联合概率分布、边缘概率分布

定义 3.1.2　如果二维随机向量 (X, Y) 所有可能取值至多为可数个，则称 (X, Y) 为**二维离散型随机向量(discrete random vector)**。

显然，如果 (X, Y) 为二维离散型随机向量，它的分量 X 和 Y 都是离散型随机变量，反之亦然。与离散型随机变量相似，通常我们用一系列等式来描述离散型随机向量的概率分布。

设(X, Y)的所有可能取值为(x_i, y_j)记

$$P(X=x_i, Y=y_j)=p_{ij} \quad (i=1, 2, \cdots, j=1, 2, \cdots) \tag{3.1.1}$$

称式(3.1.1)为**离散型随机向量(X, Y)的联合概率分布(joint probability distribution)**。

显然,式(3.1.1)中的 p_{ij} 具有如下性质:

(1) $p_{ij} \geqslant 0$;

(2) $\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} p_{ij}=1$;

(3) $P((X, Y) \in D)=\sum_{(x_i, y_j)\in D} p_{ij}$

为直观起见,有时用概率分布表(见表 3.1.1)来表示二维离散型随机向量(X, Y)的联合概率分布。

表 3.1.1

Y / X	y_1	y_2	$\cdots$	y_j	$\cdots$
x_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$
x_2	p_{21}	p_{22}	$\cdots$	p_{2j}	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$
x_i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$

在随机向量(X, Y)中,随机变量 X 的分布称为(X, Y)关于 X 的**边缘概率分布(marginal probability distrbution)**,随机变量 Y 的分布称为(X, Y)关于 Y 的**边缘概率分布**。

对于二维离散型随机向量(X, Y),如果已知(X, Y)的联合概率分布,容易求得相应的边缘概率分布。设(X, Y)的联合概率为

$$P(X=x_i, Y=y_j)=p_{ij}, (i=1, 2, \cdots, j=1, 2, \cdots)$$

则(X, Y)关于 X 和 Y 的边缘概率分布分别为

$$p_{i\cdot}=P(X=x_i)=\sum_j p_{ij}, i=1, 2, \cdots$$

$$p_{\cdot j}=P(Y=y_j)=\sum_i p_{ij}, j=1, 2, \cdots$$

例 3.1.1 有一批产品共 12 件,其中 9 件一等品,3 件二等品,现从中连续不

放回地抽取两次，每次 1 件。记 X 为第一次取到的一等品数，Y 为第二次取到的一等品数。求

① (X, Y)的联合概率分布；② (X, Y)关于 X, Y 的边缘概率分布。

解　依题知(X, Y)所有可能的取值为$(0, 0)$, $(0, 1)$, $(1, 0)$, $(1, 1)$。因为

$$P(X=0, Y=0)=P(X=0)\cdot P(Y=0\mid X=0)=\frac{C_3^1}{C_{12}^1}\cdot\frac{C_2^1}{C_{11}^1}=\frac{3}{12}\times\frac{2}{11}=\frac{1}{22};$$

$$P(X=0, Y=1)=P(X=0)\cdot P(Y=1\mid X=0)=\frac{C_3^1}{C_{12}^1}\cdot\frac{C_9^1}{C_{11}^1}=\frac{3}{12}\times\frac{9}{11}=\frac{9}{44};$$

$$P(X=1, Y=0)=P(X=1)\cdot P(Y=0\mid X=1)=\frac{C_9^1}{C_{12}^1}\cdot\frac{C_3^1}{C_{11}^1}=\frac{9}{12}\times\frac{3}{11}=\frac{9}{44};$$

$$P(X=1, Y=1)=P(X=1)\cdot P(Y=1\mid X=1)=\frac{C_9^1}{C_{12}^1}\cdot\frac{C_8^1}{C_{11}^1}=\frac{9}{12}\times\frac{8}{11}=\frac{6}{11};$$

所以(X, Y)的联合概率分布表为

X \ Y	0	1
0	$\frac{1}{22}$	$\frac{9}{44}$
1	$\frac{9}{44}$	$\frac{6}{11}$

由上表可得

$$P(X=0)=P(X=0, Y=0)+P(X=0, Y=1)=\frac{1}{22}+\frac{9}{44}=\frac{1}{4};$$

$$P(X=1)=P(X=1, Y=0)+P(X=1, Y=1)=\frac{9}{44}+\frac{6}{11}=\frac{3}{4};$$

于是(X, Y)关于 X 的边缘概率分布为

X	0	1
P	$\frac{1}{4}$	$\frac{3}{4}$

同样可得(X, Y)关于Y的边缘概率分布为

Y	0	1
P	$\frac{1}{4}$	$\frac{3}{4}$

注意 在本例中如果将"连续不放回地抽取两次"改为"连续有放回地抽取两次",其结果如何? 读者自己解答。

由以上讨论及本例都可发现,如果已知随机向量(X, Y)的联合概率分布,这时关于X和Y的边缘概率分布可由联合概率分布求出。很容易理解,总体的规律性(即联合概率分布)确定了,那么它的每个分量的规律性也就确定了。善于思考的读者可能会提出这样一个问题:如果反过来已知关于X和Y的边缘概率分布,能否决定二维随机向量(X, Y)的联合概率分布呢? 请看下面的例子。

例 3.1.2 下面是二维随机向量(X, Y)的联合概率分布

X \ Y	0	1
0	$\frac{1}{8}$	$\frac{1}{8}$
1	a	$\frac{5}{8}$

求 ①常数a;②(X, Y)关于X, Y的边缘概率分布。

解 ① 由离散型随机向量(X, Y)联合概率分布的性质知

$$\frac{1}{8}+\frac{1}{8}+\frac{5}{8}+a=1$$

所以$a=\frac{1}{8}$。

② 根据上面联合概率分布很容易求得关于X, Y的边缘概率分布分别为

X	0	1
P	$\frac{1}{4}$	$\frac{3}{4}$

Y	0	1
P	$\frac{1}{4}$	$\frac{3}{4}$

比较例 3.1.1 和例 3.1.2 发现,两者具有完全相同的边缘概率分布,但是它们

的联合概率分布却是不同的。由此可知,边缘概率分布并不能唯一地确定联合概率分布。事实上,二维离散型随机向量(X, Y)的联合概率分布不仅包含了边缘概率分布的内容,而且还包含X与Y之间相互关系的内容。在本章第3.2节及第4章我们将进一步讨论二维离散型随机向量(X, Y)中X与Y之间的关系:独立性和不相关性。

例 3.1.3　电子显示牌上有两排灯泡,第一排有2个,第二排3个,令X,Y分别表示在规定时间内第一排和第二排烧坏的灯泡数,根据以往的经验知(X, Y)的联合概率分布如下表

X \ Y	0	1	2	3
0	0.02	0.06	0.08	0.11
1	0.03	0.05	0.07	0.10
2	0.04	0.06	0.16	0.22

求　①第一排烧坏的灯泡数不超过1个的概率;

②第一排与第二排烧坏的灯泡数相等的概率;

③第一排烧坏的灯泡数不超过第二排烧坏的灯泡数的概率。

解　由联合概率分布得

① 所求事件的概率为

$$\begin{aligned}P(X \leqslant 1) &= \sum_{i=0}^{3} P(X=0, Y=i) + \sum_{i=0}^{3} P(X=1, Y=i) \\ &= 0.02+0.06+0.08+0.11+0.03+0.05+0.07+0.10 \\ &= 0.52\end{aligned}$$

② 所求事件的概率为

$$P(X=Y) = \sum_{i=0}^{2} P(X=i, Y=i) = 0.02+0.05+0.16 = 0.23$$

③所求事件的概率为

$$\begin{aligned}P(X \leqslant Y) &= 1 - P(X > Y) \\ &= 1 - P(X=1, Y=0) - P(X=2, Y=0) - P(X=2, Y=1) \\ &= 1 - 0.03 - 0.04 - 0.06 = 0.87\end{aligned}$$

由例3.1.3可知,求与离散型随机向量(X, Y)有关的某随机事件的概率,就是根据(X, Y)的联合概率分布,将此事件中所包含的(X, Y)取所有可能值的概率相加即可。

3. 二维连续型随机向量及其联合概率密度、边缘概率密度

定义 3.1.3　对二维随机向量(X, Y)，如果存在非负函数，使得对平面上任意矩形区域 $D=\{(X, Y)\mid a<X<b,\ c<Y<d\}$，都有

$$P\{(X, Y)\in D\}=\iint_D f(x, y)\mathrm{d}x\mathrm{d}y$$

则称(X, Y)为**二维连续型随机向量(continuous random vector)**，称非负函数 $f(x, y)$为随机变量 X 与 Y 的**联合概率密度(joint probability density)**，记为 $(X, Y)\sim f(x, y)$。

由定义知，作为 X 和 Y 的联合概率密度 $f(x, y)$具有如下性质：

(1) $f(x, y)\geqslant 0$；

(2)$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}f(x, y)\mathrm{d}x\mathrm{d}y=1$；

(3)$P\{(X, Y)\in D\}=\iint_D f(x, y)\mathrm{d}x\mathrm{d}y$。

如果(X, Y)是二维连续型随机向量，它的两个分量 X 和 Y 也分别是连续型随机变量，则称 X 的概率密度函数 $f_X(x)$为(X, Y)关于 X 的**边缘概率密度(marginal probability density)**；同样称 Y 的概率密度 $f_Y(y)$为(X, Y)关于 Y 的**边缘概率密度**。

设$(X, Y)\sim f(x, y)$，则有

$$f_X(x)=f_1(x)=\int_{-\infty}^{+\infty}f(x, y)\mathrm{d}y$$

$$f_Y(y)=f_2(y)=\int_{-\infty}^{+\infty}f(x, y)\mathrm{d}x$$

注意　与二维离散型随机向量不同，二维连续型随机向量不能简单地定义为："各分量都是一维连续型随机变量的随机向量"。例如：设随机变量 X 服从区间$[0, 1]$上的均匀分布，$Y=X$，显然随机向量(X, Y)的两个分量 X 和 Y 都是连续型的。但是由于(X, Y)的取值只能在图 3.1.1 所示的正方形的对角线上，因而对任何非负函数 $f(x,y)$，在 $\mathbf{R}^2$ 上的积分等于其在对角线上的积分，恒等于 0，即随机向量(X, Y)的联合分布密度不存在，不是连续型随机向量。

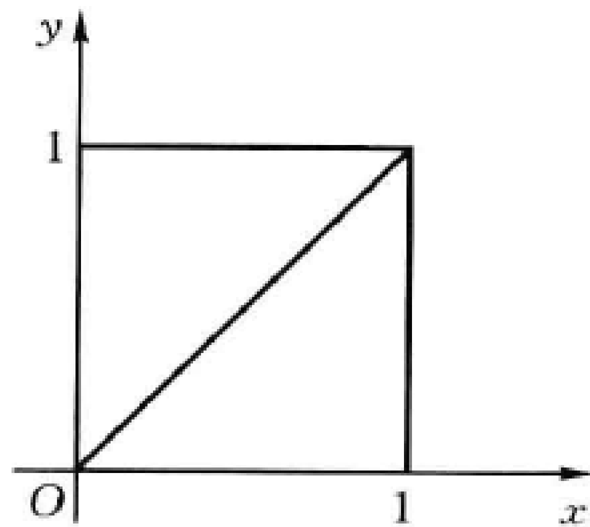


图 3.1.1

下面介绍两种常见的二维连续型随机向量。

(1)二维均匀分布

若二维随机向量(X, Y)的联合概率密度为

$$f(x, y)=\begin{cases}\dfrac{1}{S_D}, & (x, y)\in D\\ 0, & \text{其它}\end{cases}$$

其中 D 为平面上的一个可度量的有界区域，S_D 是区域 D 的面积，则称(X, Y)服从区域 D 上的均匀分布。

对于 D 中的任意可度量子区域 G，有

$$P\{(x, y)\in G\}=\iint_G f(x, y)\mathrm{d}x\mathrm{d}y=\iint_G \frac{1}{S_D}\mathrm{d}x\mathrm{d}y=\frac{S_G}{S_D}$$

其中 S_G 为 G 的面积。上式表明：二维随机向量(X, Y)落入区域 G 的概率与 G 的面积成正比，而与 G 在 D 中的位置和形状无关。由此可知，“均匀”分布的含意就是“等可能”的意思。

例 3.1.4　设二维随机向量(X, Y)在区域 $D=\{(x, y)\mid 0<x<1, |y|<x\}$ 内服从均匀分布，求① (X, Y)的联合概率密度；② (X, Y)的边缘概率密度；③ $P(X<\frac{1}{2})$。

解　① 区域 D 如图 3.1.2 所示阴影部分，其面积为

$$S_D=\frac{1}{2}\times 1\times 2=1$$

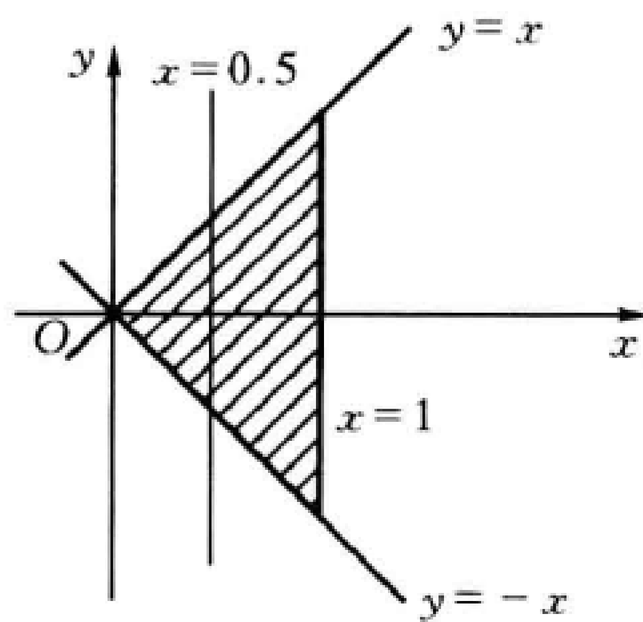


图 3.1.2

所以(X, Y)的联合概率密度为

$$f(x, y)=\begin{cases}1, & 0<x<1, |y|<x\\ 0, & \text{其它}\end{cases}$$

② 当 $0<x<1$ 时，有

$$f_1(x)=\int_{-\infty}^{+\infty}f(x, y)\mathrm{d}y=\int_{-x}^{x}\mathrm{d}y=2x$$

所以(X, Y)关于 X 的边缘概率密度为

$$f_1(x)=\begin{cases}2x, & 0<x<1\\ 0, & \text{其它}\end{cases}$$

同理，当 $-1<y<0$ 时，有

$$f_2(y)=\int_{-\infty}^{+\infty}f(x, y)\mathrm{d}x=\int_{-y}^{1}\mathrm{d}x=1+y$$

当 $0\leqslant y<1$ 时，有

$$f_2(y)=\int_{-\infty}^{+\infty}f(x, y)\mathrm{d}x=\int_{y}^{1}\mathrm{d}x=1-y$$

所以(X, Y)关于 Y 的边缘概率密度为

$$f_2(y)=\begin{cases}1+y, & -1<y<0\\ 1-y, & 0\leqslant y<1\\ 0, & \text{其它}\end{cases}$$

③所求随机事件的概率即等于(X, Y)落入区域 $G=\{(X, y)\mid 0<x<\frac{1}{2}, |y|<x\}$的概率，而 G 的面积 $S_G=\frac{1}{2}\times\frac{1}{2}\times 1=\frac{1}{4}$，由均匀分布的性质知

$$P(X<\frac{1}{2})=\frac{S_G}{S_D}=\frac{\frac{1}{4}}{1}=0.25$$

当然，利用 $f_1(x)$ 也可求得 $P(X<\frac{1}{2})=\int_0^{\frac{1}{2}}2x\mathrm{d}x=0.25$。

本例也说明一个二维均匀分布的边缘概率分布未必是均匀分布。

(2)二维正态分布

若二维随机向量(X, Y)的联合概率密度为

$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\right)\right\}$$

其中$-\infty<\mu_1, \mu_2<+\infty$，$\sigma_1^2>0$，$\sigma_2^2>0$，$|\rho|<1$，则称$(X, Y)$服从参数为 μ_1，μ_2，σ_1^2，σ_2^2，ρ 的二维正态分布，记作$(X, Y)\sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$。

二维正态分布是最重要的一种连续型随机向量的分布，其联合概率密度的图形好像一个椭球切面的钟，倒扣在 xOy 平面上，中心在(μ_1, μ_2)点。

下图 3.1.3 是利用 Mathematica 软件包画出的二维正态分布 $N(0, 0, 1, 1, 0)$ 的图像。

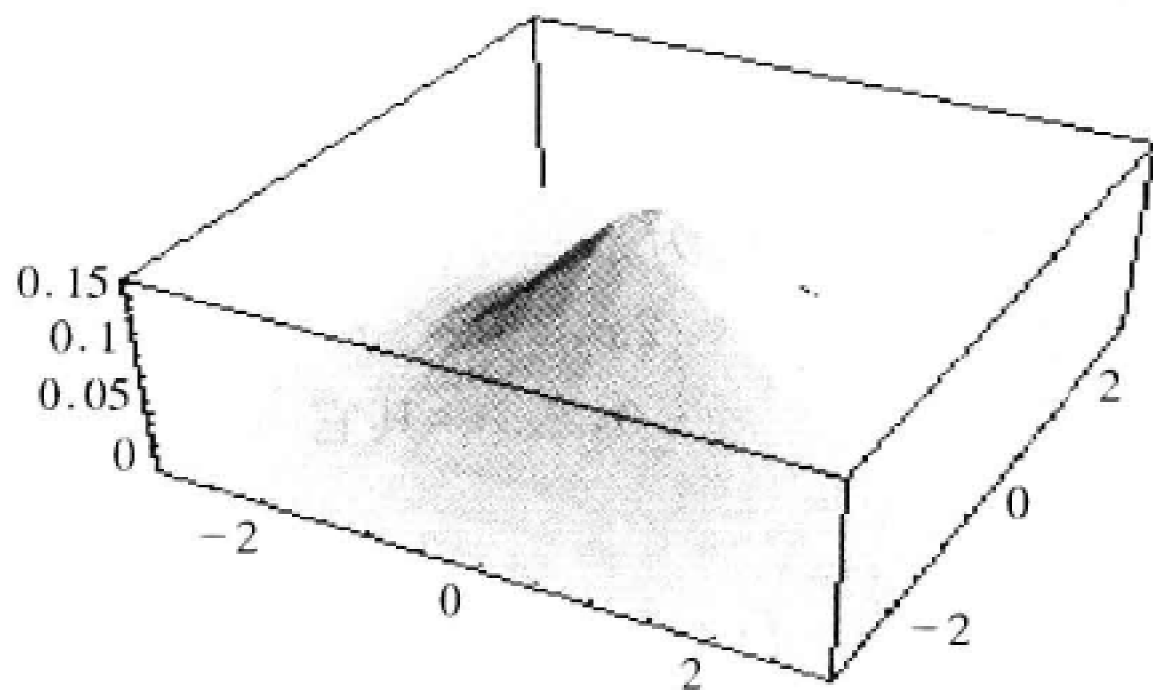
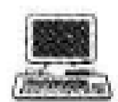


图 3.1.3

 **二维正态曲面演示实验**

显然 $f(x,y)\geqslant 0$,利用二重积分的变量代换,读者可自己验证

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,\ y)\mathrm{d}x\mathrm{d}y = 1$$

例 3.1.5　二维随机向量$(X,\ Y)\sim N(\mu_1,\ \mu_2,\ \sigma_1^2,\ \sigma_2^2,\ \rho)$,求$(X,\ Y)$的边缘概率密度。

解　由边缘概率密度的定义知

$$f_1(x) = \int_{-\infty}^{+\infty} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right)\right\}\mathrm{d}y$$

$$\xlongequal{t=\frac{y-\mu_2}{\sigma_2}} \int_{-\infty}^{+\infty} \frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho t\frac{x-\mu_1}{\sigma_1} + t^2\right)\right\}\mathrm{d}t$$

$$= \int_{-\infty}^{+\infty} \frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\left(t-\rho\frac{x-\mu_1}{\sigma_1}\right)^2 - (1-\rho^2)\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right)\right\}\mathrm{d}t$$

$$= \frac{1}{2\pi\sigma_1}\exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right]\cdot\int_{-\infty}^{+\infty} \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left(t-\rho\frac{x-\mu_1}{\sigma_1}\right)^2\right\}\mathrm{d}t$$

上式等号右边积分号下关于 t 的函数为正态分布 $N(\frac{\rho(x-\mu_1)}{\sigma_1},\ 1-\rho^2)$的概率密度函数,故其积分值等于 1,因此$(X,Y)$关于 X 的边缘概率密度为

$$f_1(x) = \frac{1}{\sqrt{2\pi}\sigma_1}\mathrm{e}^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}$$

同样可得$(X,\ Y)$关于 Y 的边缘概率密度为

$$f_2(y) = \frac{1}{\sqrt{2\pi}\sigma_2}\mathrm{e}^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}$$

此例说明服从二维正态分布 $N(\mu_1,\ \mu_2,\ \sigma_1^2,\ \sigma_2^2,\ \rho)$的随机向量$(X,\ Y)$的边缘概率分布是一维正态分布,且关于 X 的边缘概率分布为 $N(\mu_1,\ \sigma_1^2)$,关于 Y 的边缘概率分布为 $N(\mu_2,\ \sigma_2^2)$。由此可知随机向量的联合概率密度完全决定了它的边缘概率密度,但是反过来不一定成立。即边缘概率分布不足以决定随机向量的联合概率分布。例如下面两个二维正态分布

$$N(0,\ 0,\ 1,\ 1,\ \frac{1}{4}) \text{ 和 } N(0,\ 0,\ 1,\ 1,\ -\frac{1}{4})$$

它们的任一边缘概率分布都是标准正态分布 $N(0,\ 1)$,但它们却是不同的二维正态分布,因为 ρ 的值不同。对于这种现象,可以这样解释:边缘概率分布只考虑了单个随机变量的分布情况,未涉及它们之间的关系,而这个信息包含在联合概率分

布之内。就本例来说，在第4章第4.3节将证明：ρ参数恰好反映了两个分量X与Y之间的某种关系。

例3.1.6 设二维随机向量(X, Y)的联合概率密度为

$$f(x, y)=\frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}}(1+xy),\ -\infty<x,\ y<+\infty$$

求(X, Y)关于X、Y的边缘概率密度。

解
$$\begin{aligned}f_X(x)&=\int_{-\infty}^{+\infty}f(x, y)\mathrm{d}y=\frac{1}{2\pi}\int_{-\infty}^{+\infty}e^{-\frac{x^2+y^2}{2}}(1+xy)\mathrm{d}y\\&=\frac{1}{2\pi}\left(\int_{-\infty}^{+\infty}e^{-\frac{x^2+y^2}{2}}\mathrm{d}y+\int_{-\infty}^{+\infty}e^{-\frac{x^2+y^2}{2}}\cdot xy\mathrm{d}y\right)\\&=\frac{1}{2\pi}e^{-\frac{x^2}{2}}\left(\int_{-\infty}^{+\infty}e^{-\frac{y^2}{2}}\mathrm{d}y+x\int_{-\infty}^{+\infty}e^{-\frac{y^2}{2}}\cdot y\mathrm{d}y\right)\\&=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\cdot\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{y^2}{2}}\mathrm{d}y\\&=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\end{aligned}$$

同理可求得

$$f_Y(y)=\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}$$

在这个例子中，随机变量X与Y均服从标准正态分布$N(0, 1)$，但是(X, Y)却不服从二维正态分布，由此说明：边缘概率分布为正态分布的二维随机向量不一定是二维正态分布。

例3.1.7 设二维随机向量(X, Y)的联合概率密度为

$$f(x, y)=\begin{cases}e^{-(x+y)}, & x>0,\ y>0\\ 0, & \text{其它}\end{cases}$$

试求：① $P(0<X<1,\ 0<Y<1)$；②(X, Y)关于X、Y的边缘概率密度。

解 ① $$P(0<X<1,\ 0<Y<1)=\int_0^1\int_0^1 e^{-(x+y)}\mathrm{d}x\mathrm{d}y=\int_0^1 e^{-x}\mathrm{d}x\cdot\int_0^1 e^{-y}\mathrm{d}y=(1-e^{-1})^2$$

② 当$x>0$时，有

$$f_X(x)=\int_{-\infty}^{+\infty}e^{-(x+y)}\mathrm{d}y=e^{-x}\int_0^1 e^{-y}\mathrm{d}y=e^{-x};$$

当$x\leqslant 0$时，有$f_X(x)=0$。

所以(X, Y)关于X的边缘概率密度为

$$f_X(x)=\begin{cases}e^{-x}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

同样可求得(X, Y)关于Y的边缘概率密度为

$$f_Y(y)=\begin{cases}e^{-y}, & y>0\\ 0, & y\leqslant 0\end{cases}$$

例 3.1.8　设二维随机向量(X, Y)的联合概率密度为

$$f(x, y)=\begin{cases}C(R-\sqrt{x^2+y^2}), & x^2+y^2\geqslant R^2\\ 0, & \text{其它}\end{cases}$$

试求:① 常数C;② 当$R=2$时,(X, Y)在以原点为圆心,$r=1$为半径的圆域内取值的概率。

解　① $1=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}f(x, y)\mathrm{d}x\mathrm{d}y=\iint\limits_{x^2+y^2\leqslant R^2}C(R-\sqrt{x^2+y^2})\mathrm{d}x\mathrm{d}y$

$$=\int_0^{2\pi}\int_0^R C(R-\rho)\rho\mathrm{d}\rho\mathrm{d}\theta$$

$$=\frac{\pi}{3}CR^3$$

所以$C=\frac{3}{\pi R^3}$。

② 当$R=2$时,有

$$P(X^2+Y^2\leqslant 1)=\iint\limits_{x^2+y^2\leqslant 1}\frac{3}{8\pi}(R-\sqrt{x^2+y^2})$$

$$=\frac{3}{8\pi}\int_0^{2\pi}\mathrm{d}\theta\int_0^1(2-\rho)\mathrm{d}\rho$$

$$=\frac{1}{2}$$

例 3.1.9　设二维随机向量(X, Y)的联合概率密度为

$$f(x, y)=\begin{cases}4xy, & 0<x, y<1\\ 0, & \text{其它}\end{cases}$$

试求:① $P(X>Y)$; (2)$P(X=Y)$。

解　①积分区域如图 3.1.4 所示。

$$P(X>Y)=\int_0^1\mathrm{d}x\int_0^x 4xy\mathrm{d}y$$

$$=\int_0^1 2x^3\mathrm{d}x=\frac{1}{2}$$

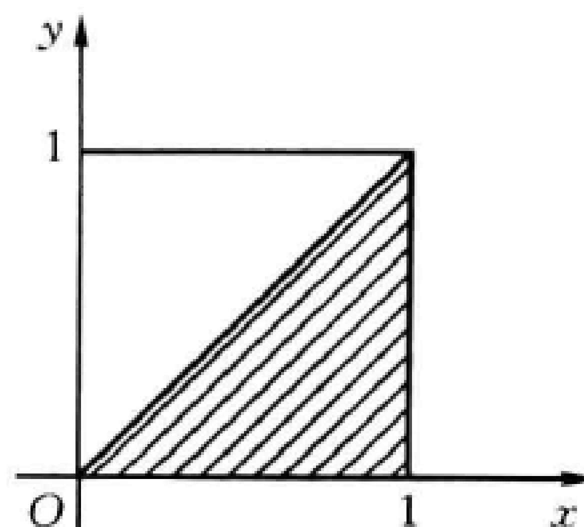


图 3.1.4

② 类似于(1)可得

$$P(X<Y)=\frac{1}{2}$$

所以

$$P(X=Y)=1-P(X>Y)-P(X<Y)=0$$

这里 $P(X=Y)=0$ 可以这样理解：由于一维连续型随机变量在一点取值的概率为 0，所以类似地有：二维连续型随机向量在一条线取值的概率为 0。

练习 3.1

1. 现有 10 件产品，其中 6 件正品，4 件次品。从中随机抽取 2 次，每次抽取 1 件，定义两个随机变量 X,Y 如下：

$$X=\begin{cases}1, & \text{第 1 次抽到正品}\\ 0, & \text{第 1 次抽到次品}\end{cases} \qquad Y=\begin{cases}1, & \text{第 2 次抽到正品}\\ 0, & \text{第 2 次抽到次品}\end{cases}$$

试就下面两种情况求 (X,Y) 的联合概率分布和边缘概率分布。

(1) 第 1 次抽取后放回；(2) 第 1 次抽取后不放回。

2. 已知 10 件产品中有 5 件一级品，2 件废品。现从这批产品中任意抽取 3 次，记其中的一级品数与废品数分别为 X,Y，求 (X,Y) 的联合概率分布和边缘概率分布。

3. 已知随机变量 X,Y 的概率分布分别为

X	-1	0	1
P	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Y	0	1
P	$\frac{1}{2}$	$\frac{1}{2}$

且 $P(XY=0)=1$，求

(1) X 和 Y 的联合概率分布；

(2) $P(X=Y)$。

4. 设二维随机向量 (X, Y) 的联合概率密度为

$$f(x, y)=\begin{cases}Ae^{-(x+2y)}, & x>0,\ y>0\\ 0, & \text{其它}\end{cases}$$

试求：(1) 常数 A；

(2) (X, Y) 关于 X, Y 的边缘概率密度；

(3) $P(0<X\leqslant 2,\ 0<Y\leqslant 3)$；

(4) $P(X+2Y\leqslant 1)$；

(5) $P(X<Y)$。

5. 设二维随机向量(X, Y)的联合概率密度为

$$f(x, y)=\begin{cases} x^2+\dfrac{1}{3}xy, & 0\leqslant x\leqslant 1,\ 0\leqslant y\leqslant 2 \\ 0, & \text{其它} \end{cases}$$

试求:(1)(X, Y)关于 X, Y 的边缘概率密度;

(2)$P\left(X<\dfrac{1}{2}\,|\,Y<\dfrac{1}{2}\right)$。

6. 设二维随机向量(X, Y)的联合概率密度为

$$f(x,y)=\begin{cases} e^{-y}, & x>0,\ y>x \\ 0, & \text{其它} \end{cases}$$

试求:(1)(X, Y)关于 X, Y 的边缘概率密度;

(2)$P(X>2, Y<4)$。

7. 某公司经理和他的秘书定于本周星期日中午 12 点至下午 1 点在办公室会面,并约定先到者等 20 分钟后即可离去,试求二人能会面的概率。

3.2　随机向量的联合分布函数

1. 联合分布函数

定义 3.2.1　设(X, Y)二维随机向量,称 $\mathbf{R}^2$ 上的二元函数

$$F(x,y)=P(X\leqslant x, Y\leqslant y)$$

为(X, Y)的**联合分布函数(joint distribution function)**,简称**联合分布**。

如果(X, Y)为离散型随机向量,其联合概率分布为

$$P(X=x_i, Y=y_j)=p_{ij}\quad (i=1, 2, \cdots, j=1, 2, \cdots)$$

则相应的联合分布函数为

$$F(x, y)=\sum_{x_i\leqslant x}\sum_{y_j\leqslant y}p_{ij}$$

如果(X, Y)为连续型随机向量,其联合概率密度为 $f(x, y)$,则相应的联合分布函数为

$$F(x, y)=\int_{-\infty}^{x}\int_{-\infty}^{y}f(s, t)\mathrm{d}s\,\mathrm{d}t$$

在这里可以看出,分布函数 $F(x, y)$的值恰好是随机向量(X, Y)取值于区域$\{-\infty<X\leqslant x, -\infty<Y\leqslant y\}$内的概率,它是随机变量 X 的分布函数 $F(x)$的推广,因此对区域$\{x_1<X\leqslant x_2, y_1<Y\leqslant y_2\}$,有

$$P(x_1<X\leqslant x_2, y_1<Y\leqslant y_2)=F(x_2, y_2)-F(x_2, y_1)-F(x_1, x_2)+F(x_1, y_1)$$

另外 $F(x, y)=P(X\leqslant x, Y\leqslant y)$ 可以这样理解

$$F(x, y)=P\{(X\leqslant x)\cap(Y\leqslant y)\}$$

因此 $F(x, y)$ 有如下性质：

(1) $0\leqslant F(x, y)\leqslant 1$；

(2) $F(+\infty, +\infty)=1$，$F(-\infty, -\infty)=F(-\infty, 0)=F(0, -\infty)=0$。

定义 3.2.2 设(X, Y)二维随机向量的联合分布函数为 $F(x, y)$，则称

$$F_X(x)=F(x, +\infty)=P(X\leqslant x, Y<+\infty)=P(X\leqslant x),$$

$$F_Y(x)=F(+\infty, y)=P(X<+\infty, Y\leqslant y)=P(Y\leqslant y)$$

分别为 X 与 Y 的**边缘分布函数(marginal distribution function)**。

对于 $f(x, y)$ 的连续点，有

$$\frac{\partial^2 F(x, y)}{\partial x\partial y}=f(x, y)$$

2. 随机变量的相互独立性

在第 1 章我们研究了随机事件的独立性，而随机变量不过是随机试验结果定量化的描述，因此很容易得到两个随机变量 X 与 Y 的独立性，即 X 与 Y 分别对应的所有随机事件相互独立。

定义 3.2.3 设(X, Y)二维随机向量，如果对于任意的 $a<b$，$c<d$ 均有

$$P(a<X<b, c<Y<d)=P(a<X<b)\cdot P(c<Y<d)$$

则称**随机变量 X 与 Y 相互独立**。

在具体实际应用过程中，用下面的方法来判断随机变量的独立性有时更为方便。

定理 3.2.1 (1)若(X, Y)为离散型随机向量，联合概率分布为

$$P(X=x_i, Y=y_j)=p_{ij}\quad (i=1, 2, \cdots, j=1, 2, \cdots)$$

则 X 与 Y 相互独立的充分必要条件是：对任何 $i, j=1, 2, \cdots$ 都有

$$P(X=x_i, Y=y_j)=P(X=x_i)\cdot P(Y=y_j)$$

即

$$p_{ij}=p_{i\cdot}\cdot p_{\cdot j}$$

(2)若(X, Y)为连续型随机向量，联合概率密度为 $f(x, y)$，则 X 与 Y 相互独立的充分必要条件是其联合概率密度等于两个边缘概率密度的乘积，即对于任意 $(x, y)\in\mathbf{R}^2$，有

$$f(x, y)=f_1(x)\cdot f_2(y)$$

证 下面只证(2)，对于(1)留给读者自己证明。

必要性

因 X 与 Y 相互独立,故依定义 3.2.3 知:对于任意的 $a<b,c<d$ 均有

$$P(a<X<b,\ c<Y<d)=P(a<X<b)\cdot P(c<Y<d)$$

而

$$P(a<X<b)\cdot P(c<Y<d)=\int_a^b f_1(x)\mathrm{d}x\cdot\int_c^d f_2(y)\mathrm{d}y$$
$$=\int_a^b\int_c^d f_1(x)f_2(y)\mathrm{d}y\mathrm{d}x$$

所以

$$P(a<X<b,\ c<Y<d)=\int_a^b\int_c^d f_1(x)f_2(y)\mathrm{d}y\mathrm{d}x$$

这里 $f_1(x)f_2(y)\geqslant 0$,由联合概率密度的定义知 $f_1(x)f_2(y)$ 为 (X,Y) 的联合概率密度,定理结论成立。

充分性

由于 $f(x,y),f_1(x)\cdot f_2(y)$ 所以

$$P(a<X<b,\ c<Y<d)=\int_a^b\int_c^d f(x,y)\mathrm{d}y\mathrm{d}x$$
$$=\int_a^b\int_c^d f_1(x)f_2(y)\mathrm{d}y\mathrm{d}x$$
$$=\int_a^b f_1(x)\mathrm{d}x\cdot\int_c^d f_2(y)\mathrm{d}y$$
$$=P(a<X<b)\cdot P(c<Y<d)$$

由独立性的定义知 X 与 Y 相互独立。

用此定理可判别例 3.1.6 中的随机变量 X 与 Y 不独立,而例 3.1.7 中的随机变量 X 与 Y 相互独立。

定理 3.2.2　若随机变量 X 与 Y 相互独立,则连续函数 $g(X)$ 与 $h(Y)$ 也相互独立。特别地,两个相互独立的随机变量 X 与 Y 的线性函数 $g(X)$ 与 $h(Y)$ 仍独立。

此定理的证明已超出本书的范围,但此结论在今后学习过程中会经常用到。

例 3.2.1　设 A,B 为随机事件,且 $P(A)=\frac{1}{6}$, $P(B)>0$, $P(B|A)=\frac{1}{3}$, $P(A|B)=\frac{1}{6}$,令

$$X=\begin{cases}1, & \text{若 } A \text{ 发生}\\ 2, & \text{若 } A \text{ 不发生}\end{cases}\qquad X=\begin{cases}1, & \text{若 } B \text{ 发生}\\ 0, & \text{若 } B \text{ 不发生}\end{cases}$$

求 (X,Y) 的联合概率分布。

解 依题因

$$P(B)=\frac{P(A)P(B|A)}{P(A|B)}=\frac{\frac{1}{6}\times\frac{1}{3}}{\frac{1}{6}}=\frac{1}{3}$$

所以

$$P(AB)=\frac{1}{6}\times\frac{1}{3}=P(A)\cdot P(B)$$

故 A 与 B 相互独立。从而$\overline{A}$与 B，A 与$\overline{B}$，$\overline{A}$与$\overline{B}$均相互独立。即随机变量 X 与 Y 相互独立。又由于

X	1	2
P	$\frac{1}{6}$	$\frac{5}{6}$

Y	0	1
P	$\frac{1}{3}$	$\frac{2}{3}$

所以由定理 3.2.1 求得(X, Y)的联合概率分布为

X \ Y	0	1
1	$\frac{1}{18}$	$\frac{1}{9}$
2	$\frac{5}{18}$	$\frac{5}{9}$

例 3.2.2 设随机变量 X 与 Y 相互独立，下表给出了二维随机向量(X, Y)的联合概率分布及关于 X 与 Y 的边缘概率分布中的部分值，试将其余的值填入表中的空白处。

X \ Y	y_1	y_2	y_3	$p_{i\cdot}$
x_1		$\frac{1}{8}$		
x_2	$\frac{1}{8}$			
$p_{\cdot j}$	$\frac{1}{6}$			1

解 依题根据上表

由 $\frac{1}{6}=p_{\cdot 1}=p_{11}+p_{21}=p_{11}+\frac{1}{8}$，得 $p_{11}=\frac{1}{24}$。

于是由 X 与 Y 的独立性知

$p_{1\cdot}=\frac{p_{11}}{p_{\cdot 1}}=\frac{1}{4}$，　　$p_{2\cdot}=\frac{p_{21}}{p_{\cdot 1}}=\frac{3}{4}$，

$p_{\cdot 2}=\frac{p_{12}}{p_{\cdot 1}}=\frac{1}{2}$，　　$p_{\cdot 3}=1-\frac{1}{6}-\frac{1}{2}=\frac{1}{3}$，

$p_{13}=p_{1\cdot}\cdot p_{\cdot 3}=\frac{1}{4}\times\frac{1}{3}=\frac{1}{2}$，　　$p_{23}=p_{2\cdot}\cdot p_{\cdot 3}=\frac{3}{4}\times\frac{1}{3}=\frac{1}{4}$

所以(X, Y)的联合概率分布为

X \ Y	y_1	y_2	y_3	$p_{i\cdot}$
x_1	$\frac{1}{24}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{4}$
x_2	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{3}{4}$
$p_{\cdot j}$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$	1

在数理统计中，经常考虑 n 个随机变量独立性，上面这些概念和定理均可推广为 n 个随机变量的情形。

定义 3.2.4　设$(X_1, X_2, \cdots, X_n)$是 n 维随机向量，如果对于任意的 $a_i<b_i$ $(i=1, 2, \cdots)$，有

$$P(a_1<X_1<b_1, a_2<X_2<b_2, \cdots, a_n<X_n<b_n)$$
$$=P(a_1<X_1<b_1)\cdot P(a_2<X_2<b_2)\cdots P(a_n<X_n<b_n)$$

则称这 n 个随机变量 $X_1, X_2, \cdots, X_n$ 相互独立。

定理 3.2.1 的推广如下。

(1)n 个离散型随机变量 $X_1, X_2, \cdots, X_n$ 相互独立的充分必要条件是：对任何 $x_1, x_2, \cdots, x_n$，有

$$P(X_1=x_1, X_2=x_2, \cdots, X_n=x_n)$$
$$=P(X_1=x_1)\cdot P(X_2=x_2)\cdots P(X_n=x_n)$$

(2)n 个连续型随机变量 $X_1, X_2, \cdots, X_n$ 相互独立的充分必要条件是：它们的联合密度等于边缘概率分布的乘积，即对任何 $x_1, x_2, \cdots, x_n$，有

$$f(x_1, x_2, \cdots, x_n)=f_1(x_1)\cdot f_2(x_2)\cdots f_n(x_n)$$

由此可以证明：n 个相互独立的随机变量 X_1，X_2，…，X_n 中的任意 $m(1<m\leqslant n)$ 个随机变量 X_{i_1}，X_{i_2}，…，X_{i_m} 也相互独立。

定义 3.2.5 若随机变量序列 X_1，X_2，…，X_n，…中任意 $m(m\geqslant 2)$ 个随机变量相互独立，则称该随机变量序列 X_1，X_2，…，X_n，…是相互独立的，如果对所有的 X_i 服从相同的分布，则称 X_1，X_2，…，X_n，…为独立同分布的随机变量序列。

3. 随机向量函数的分布

在数理统计中，对某一随机现象的 n 次观察或试验数据可看作 n 个随机变量 X_1，X_2，…，X_n，在研究的过程中为了某种目的常常需将这些数据进行"加工"而得到一些新的随机变量，这些新的随机变量就是 X_1，X_2，…，X_n 的函数。例如若 X_1，X_2，…，X_n 是对某个未知量 μ 进行的 n 次观测的结果，由于测量中误差的存在，所以我们用 X_1，X_2，…，X_n 的平均值 $\overline{X}=\dfrac{X_1+X_2+\cdots+X_n}{n}$ 来估计未知量 μ，这里随机变量 $\overline{X}$ 就是 X_1，X_2，…，X_n 的函数。为此需要研究随机向量函数的分布。

例 3.2.3 设随机变量 X_1 与 X_2 相互独立，分别服从二项分布 $B(n_1, p)$ 和 $B(n_2, p)$，求 $Y=X_1+X_2$ 的概率分布。

解 依题知 Y 的可能取值为 0，1，2，…，n_1+n_2，因此对于 $k(0\leqslant k\leqslant n_1+n_2)$，由 X_1 与 X_2 独立性有

$$\begin{aligned}P(Y=k)&=\sum_{k_1+k_2=k}P(X_1=k_1, X_2=k_2)\\&=\sum_{k_1+k_2=k}C_{n_1}^{k_1}p^{k_1}(1-p)^{n_1-k_1}C_{n_2}^{k_2}p^{k_2}(1-p)^{n_2-k_2}\\&=\sum_{k_1+k_2=k}C_{n_1}^{k_1}C_{n_2}^{k_2}p^{k}(1-p)^{n_1+n_2-k}\end{aligned}$$

应用组合数公式 $\sum\limits_{k_1+k_2=k}C_{n_1}^{k_1}C_{n_2}^{k_2}=C_{n_1+n_2}^{k}$ 得

$$P(Y=k)=C_{n_1+n_2}^{k}p^{k}(1-p)^{n_1+n_2-k}$$

所以 Y 服从二项分布 $B(n_1+n_2, p)$。

事实上，我们从概率论的意义很容易理解这一结论。若 $X\sim B(n, p)$，则依二项分布的试验背景（n 重贝努里试验）知 X 表示在 n 次独立重复试验中事件 A 出现的次数，而每次试验中事件 A 出现的概率 p 不变。因此本例中 X_i 表示在 n_i 次试验中事件 A 出现的次数，而每次试验中事件 A 出现的概率均为 p，所以 $Y=X_1+X_2$ 表示在 n_1+n_2 次试验中事件 A 出现的次数，而每次试验中事件 A 出现的概率均为 p，于是可知 $Y\sim B(n_1+n_2, p)$。

本例说明两个相互独立的二项分布的随机变量的和仍然是二项分布，称为二项分布的"可加性"。类似地可证明，Poisson 分布也具有"可加性"。

下面考虑连续型随机变量和的概率密度函数。

例 3.2.4　设随机变量 X_1 与 X_2 相互独立，其概率密度函数分别为 $f_1(x_1)$ 和 $f_2(x_2)$，求 $Y=X_1+X_2$ 的概率密度函数。

解　因 X_1 与 X_2 相互独立，所以随机向量 (X_1, X_2) 的联合概率密度为 $f(x_1, x_2)=f_1(x_1)\cdot f_2(x_2)$，于是对任何 $a<b$，令 $y=x_1+x_2$，均有

$$\begin{aligned}P(a<Y<b) &= P(a<X_1+X_2<b)\\ &= \iint\limits_{a<x_1+x_2<b} f_1(x_1)\cdot f_2(x_2)\mathrm{d}x_1\mathrm{d}x_2\\ &= \int_{-\infty}^{+\infty}\left(\int_{a-x_2}^{b-x_2} f_1(x_1)\cdot f_2(x_2)\mathrm{d}x_1\right)\mathrm{d}x_2\\ &= \int_{-\infty}^{+\infty} f_2(x_2)\left(\int_a^b f_1(y-x_2)\mathrm{d}y\right)\mathrm{d}x_2\\ &= \int_a^b\left(\int_{-\infty}^{+\infty} f_1(y-x)\cdot f_2(x)\mathrm{d}y\right)\mathrm{d}x\end{aligned}$$

上式中 $\int_{-\infty}^{+\infty} f_1(y-x)\cdot f_2(x)\mathrm{d}x$ 即为 Y 的概率密度函数。即

$$f_Y(y) = \int_{-\infty}^{+\infty} f_1(y-x)\cdot f_2(x)\mathrm{d}x$$

作变换 $t=y-x$，然后把积分变量 t 再换回到 x，又可得到

$$f_Y(y) = \int_{-\infty}^{+\infty} f_1(x)\cdot f_2(y-x)\mathrm{d}x$$

例 3.2.5　设随机变量 X_1 与 X_2 相互独立，且均服从标准正态分布 $N(0, 1)$，求 $Y=X_1+X_2$ 的概率密度函数。

解　依题知 X_i 的概率密度函数为

$$f_i(x_i) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x_i^2}{2}},\ i=1,\ 2$$

因 X_1 与 X_2 相互独立，利用例 3.2.4 的结果可知 Y 的概率密度函数为

$$\begin{aligned}f_Y(y) &= \int_{-\infty}^{+\infty} f_1(y-x)\cdot f_2(x)\mathrm{d}x\\ &= \frac{1}{2\pi}\int_{-\infty}^{+\infty}\mathrm{e}^{-\frac{x^2}{2}}\cdot\mathrm{e}^{-\frac{(y-x)^2}{2}}\mathrm{d}x\\ &= \frac{1}{2\pi}\mathrm{e}^{-\frac{y^2}{4}}\int_{-\infty}^{+\infty}\mathrm{e}^{-\left(x-\frac{y}{2}\right)^2}\mathrm{d}x\end{aligned}$$

令 $\frac{t}{\sqrt{2}}=x-\frac{y}{2}$，并利用 $\int_{-\infty}^{+\infty}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t=\sqrt{2\pi}$ 即得

$$f_Y(y)=\frac{1}{2\sqrt{2}\pi}e^{-\frac{y^2}{4}}\int_{-\infty}^{+\infty}e^{-\frac{t^2}{2}}dt=\frac{1}{2\sqrt{\pi}}e^{-\frac{y^2}{4}}$$

这是正态分布 $N(0, 2)$的概率密度函数。

上述结论对于一般的正态分布也成立,即:随机变量 $X_1\sim N(\mu_1,\sigma_1^2)$, $X_2\sim N(\mu_2,\sigma_2^2)$且相互独立,则 $Y=X_1+X_2\sim N(\mu_1+\mu_2,\sigma_2^2+\sigma_2^2)$。这说明两个相互独立的正态随机变量之和仍是正态分布,且有关的参数相加。这个性质称为正态分布的"可加性"。用类似的方法将此结论推广为下面更一般的结果。

有限个相互独立的正态随机变量的线性函数仍服从正态分布。即若 $X_i\sim N(\mu_i,\sigma_i^2)$, $i=1, 2, \cdots, n$,且 $X_1, X_2, \cdots, X_n$ 相互独立,$a_1, a_2, \cdots, a_n$ 是不全为零的常数,则有

$$\sum_{i=1}^n a_iX_i\sim N(\sum_{i=1}^n a_i\mu_i,\ \sum_{i=1}^n a_i^2\sigma_i^2)$$

特别地,若 $X_1, X_2, \cdots, X_n$ 相互独立且均服从正态分布 $N(\mu,\sigma^2)$,则有

$$\overline{X}=\frac{1}{n}\sum_{i=1}^n X_i\sim N(\mu,\frac{\sigma^2}{n})$$

例 3.2.6 设二维随机向量$(X, Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$,$f(x, y)$为其联合概率密度,则 X 与 Y 相互独立的充要条件是$\rho=0$。

证 依题知

$$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}},$$

$$f_Y(y)=\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}$$

因此

$$f_X(x)\cdot f_Y(y)=\frac{1}{2\pi\sigma_1\sigma_2}\exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2+\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

充分性:若 $\rho=0$,将其代入 $f(x, y)$中知对于任意 x,y,有

$$f(x, y)=f_X(x)\cdot f_Y(y)$$

这说明 X 与 Y 相互独立。

必要性:若 X 与 Y 相互独立,则有

$$f(x, y)=f_X(x)\cdot f_Y(y),\quad -\infty<x, y<+\infty$$

特别取 $x=\mu_1$, $y=\mu_2$ 得

$$f(\mu_1,\mu_2)=f_X(\mu_1)\cdot f_Y(\mu_2)$$

即

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}=\frac{1}{2\pi\sigma_1\sigma_2}$$

从而知 $\rho=0$。

下面从分布函数与概率密度函数的关系出发，介绍另一种求随机变量函数的概率密度函数的重要方法——分布函数法。

设随机向量 (X, Y) 的联合概率密度为 $f(x, y)$，记 $Z=g(X, Y)=X+Y$，则随机变量 Z 的分布函数为

$$F(z)=P(Z\leqslant z)=P(g(X, Y)\leqslant z)$$
$$=\iint_{g(x,y)\leqslant z} f(x, y)\mathrm{d}x\mathrm{d}y$$
$$=P(X+Y\leqslant z)=\iint_{x+y\leqslant z} f(x, y)\mathrm{d}x\mathrm{d}y$$

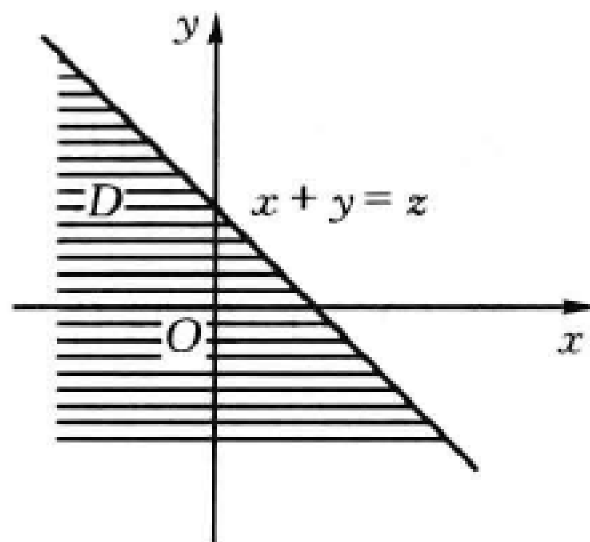


图 3.2.1

将 z 看作常数化为二次积分计算其值，求出 Z 的分布函数 $F(z)$，积分区域如图 3.2.1 阴影部分。再根据分布函数与概率密度函数的关系，对 $F(z)$ 求关于 z 的导数即可得到 Z 的概率密度函数 $f(z)=F'(z)$。

这就是利用分布函数法求随机变量函数的概率密度函数的基本步骤，这种方法很有代表性，在求一般随机变量函数（包括一元或多元）的概率密度函数时非常有效，解题时经常用到，应予以重视。

例 3.2.7　设随机变量 X 与 Y 相互独立，其概率密度函数

$$f_X(x)=\begin{cases}\dfrac{2}{\sqrt{\pi}}\mathrm{e}^{-x^2}, & x>0\\ 0, & \text{其它}\end{cases}\qquad f_Y(y)=\begin{cases}\dfrac{2}{\sqrt{\pi}}\mathrm{e}^{-y^2}, & y>0\\ 0, & \text{其它}\end{cases}$$

求 $Z=\sqrt{X^2+Y^2}$ 的概率密度函数。

解　设 (X, Y) 的联合概率密度函数为

$$f(x,y)=f_X(x)\cdot f_Y(y)=\begin{cases}\dfrac{4}{\pi}\mathrm{e}^{-(x^2+y^2)}, & x>0,\ y>0\\ 0, & \text{其它}\end{cases}$$

因此当 $z<0$ 时，$F_Z(z)=0$，从而 $f_Z(z)=0$。

当 $z\geqslant 0$ 时，有

$$F_Z(z)=\iint_{Z\leqslant z} f(x, y)\mathrm{d}x\mathrm{d}y=\frac{4}{\pi}\iint_{x^2+y^2\leqslant z^2}\mathrm{e}^{-(x^2+y^2)}\mathrm{d}x\mathrm{d}y$$
$$=\frac{4}{\pi}\int_0^{\frac{\pi}{2}}\mathrm{d}\theta\int_0^z \mathrm{e}^{-r^2}r\mathrm{d}r$$
$$=1-\mathrm{e}^{-z^2}$$

于是

$$f_Z(z)=F'_Z(z)=2z\mathrm{e}^{-z^2}$$

故 Z 的概率密度函数为

$$f_Z(z)=\begin{cases}2z\mathrm{e}^{-z^2}, & z\geqslant 0\\ 0, & z<0\end{cases}$$

例 3.2.8 设随机向量 (X, Y) 服从区域 $D=\{(x, y)\mid 1\leqslant x\leqslant 3, 1\leqslant y\leqslant 3\}$ 上的均匀分布，求 $U=|X-Y|$ 的概率密度函数。

解 (X, Y) 的联合概率密度函数为

$$f(x,y)=\begin{cases}\dfrac{1}{4}, & 1\leqslant x\leqslant 3,\ 1\leqslant y\leqslant 3\\ 0, & \text{其它}\end{cases}$$

因此当 $u\leqslant 0$ 时，$F(u)=0$；

当 $0<u<2$ 时，积分区域如图 3.2.2 阴影部分，于是有

$$\begin{aligned}F(u)&=\iint\limits_{|x-y|\leqslant u} f(x, y)\mathrm{d}x\mathrm{d}y\\&=\iint\limits_{|x-y|\leqslant u}\frac{1}{4}\mathrm{d}x\mathrm{d}y\\&=\frac{1}{4}[4-(2-u)^2]\\&=u-\frac{u^2}{4}\end{aligned}$$

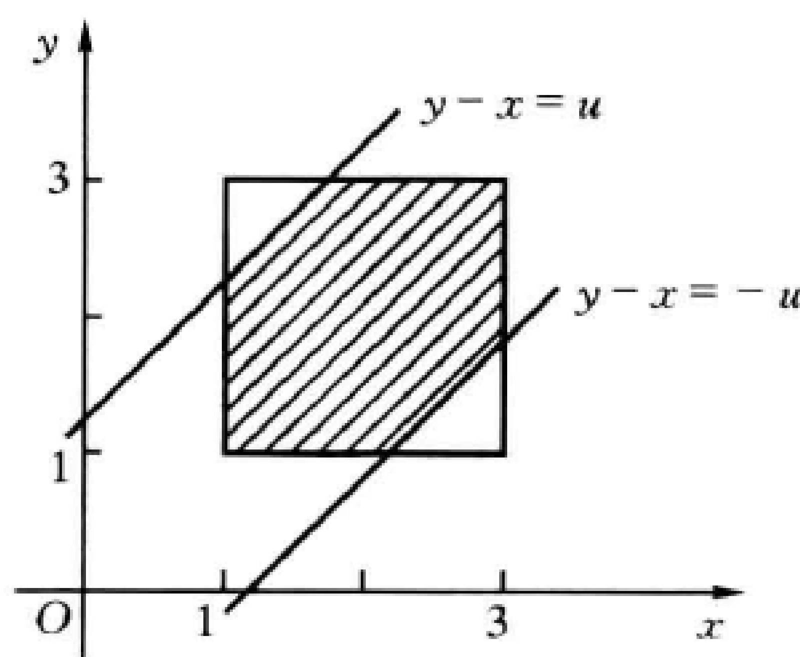


图 3.2.2

当 $u\geqslant 2$ 时，$F(u)=1$。

故 Z 的概率密度函数为

$$f_U(u)=\begin{cases}1-\dfrac{1}{2}u, & 0<u<2\\ 0, & \text{其它}\end{cases}$$

例 3.2.9 设某种型号的晶体管的寿命(以小时计)近似地服从正态分布 $N(160, 20^2)$，现从中随机取 4 只，求其中没有 1 只晶体管的寿命小于 180 小时的概率。

解 设 $X_i(i=1, 2, 3, 4)$ 表示第 i 只晶体管的寿命，于是 $X_i(i=1, 2, 3, 4)$ 相互独立，且同正态分布 $N(160, 20^2)$，其概率密度函数均为

$$f_i(x)=\frac{1}{20\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-160)^2}{2\times 20^2}}$$

因此所求概率为

$$
\begin{aligned}
& P\{\min(X_1, X_2, X_3, X_4) > 180\} \\
&= P(X_1 > 180, X_2 > 180, X_3 > 180, X_4 > 180) \\
&= P(X_1 > 180) \cdot P(X_2 > 180) \cdot P(X_3 > 180) \cdot P(X_4 > 180) \\
&= \left[\int_{80}^{+\infty} \frac{1}{20\sqrt{2\pi}} e^{-\frac{(x-160)^2}{2\times 20^2}} dx\right]^4 \\
&= \left[1 - \Phi\left(\frac{180-160}{20}\right)\right]^4 \\
&= [1 - \Phi(1)]^4 \\
&= (1 - 0.8413)^4 \\
&= 0.00634
\end{aligned}
$$

练习 3.2

1. 设随机变量 X 与 Y 相互独立同分布，且 $P(X=-1)=P(Y=-1)=\frac{1}{2}$，$P(X=1)=P(Y=1)=\frac{1}{2}$，则(　　)。

(A) $P(X=Y)=\frac{1}{2}$　　(B) $P(X=Y)=1$

(C) $P(X+Y=0)=\frac{1}{4}$　　(D) $P(XY=1)=\frac{1}{4}$

2. 设随机变量 $X_i\ (i=1, 2, 3, 4)$ 相互独立同分布，且 $P(X_i=0)=0.6$，$P(X_i=1)=0.4\ (i=1, 2, 3, 4)$，求行列式 $X=\begin{vmatrix} X_1 & X_2 \\ X_3 & X_4 \end{vmatrix}$ 的分布列。

3. 设二维随机向量 (X, Y) 服从矩形区域 $D=\{(x, y) \mid 0\leqslant x\leqslant 2, 0\leqslant y\leqslant 1\}$ 上的均匀分布，且

$$U=\begin{cases} 0, & X\leqslant Y \\ 1, & X>Y \end{cases}\qquad V=\begin{cases} 0, & X\leqslant 2Y \\ 1, & X>2Y \end{cases}$$

求 U 与 V 的联合概率分布。

4. 求练习 3.1 第 4,5,6 题中 (X, Y) 的联合分布函数。

5. 设二维随机向量 (X, Y) 的联合概率密度为

$$f(x, y)=\begin{cases} 4xy, & 0\leqslant x\leqslant 1, 0\leqslant y\leqslant 1 \\ 0, & \text{其它} \end{cases}$$

求 (X, Y) 的联合分布函数。

6. 设随机变量 X 与 Y 相互独立，其概率密度函数分别为

$$f_X(x)=\begin{cases}1, & 0\leqslant x\leqslant 1\\ 0, & \text{其它}\end{cases} \qquad f_Y(y)=\begin{cases}Ae^{-y}, & y>0\\ 0, & y\leqslant 0\end{cases}$$

求：(1)常数 A；

(2)随机变量 $Z=2X+Y$ 的概率密度函数。

7. 设(X, Y)的联合分布函数为

$$F(x, y)=A\left(B+\arctan\frac{x}{3}\right)\left(C+\arctan\frac{y}{4}\right)$$

求：(1)常数 A, B, C；

(2)(X, Y)的联合概率密度；

(3)(X, Y)的边缘分布函数和边缘概率密度；

(4) $P(X<3)$, $P(Y<4)$, $P(X<3, Y<4)$；(5) 判断 X 与 Y 的独立性。

8. 设某仪器由两个部件构成，用 X、Y 分别表示两个部件的寿命(单位：千小时)，已知(X, Y)的联合分布函数为

$$F(x, y)=\begin{cases}1-e^{-0.5x}-e^{-0.5y}+e^{-0.5(x+y)}, & x>0, y>0\\ 0, & \text{其它}\end{cases}$$

试求：(1)求(X, Y)的两个边缘分布函数；

(2)求(X, Y)联合概率密度与边缘概率密度；

(3)X 与 Y 是否独立；

(4)两个部件寿命都超过 100 小时的概率。

9. 设 X 与 Y 相互独立，且 X 服从$\lambda=3$ 的指数分布，Y 服从$\lambda=4$ 的指数分布，试求：

(1)(X, Y)联合概率密度与边缘概率密度；

(2)$P(X<1, Y<1)$；

(3)(X, Y)在 $D=\{(x,y|x>0, y>0, 3x+4y<3)\}$取值的概率。

10. 对随机变量 X,Y 有

$$P(X\geqslant 0, Y\geqslant 0)=\frac{3}{7}, \quad P(X\geqslant 0)=P(Y\geqslant 0)=\frac{4}{7}$$

求 $P\{\max(X, Y)\geqslant 0\}$, $P\{\min(X, Y)<0\}$。

11. (X, Y)的联合概率密度为

$$f(x, y)=\begin{cases}3x, & 0<x<1, 0<y<x\\ 0, & \text{其它}\end{cases}$$

求 $Z=X-Y$ 概率密度函数。

3.3 条件分布

对于二维随机向量(X, Y)而言，随机变量 X 的条件分布是指在给定 Y 取某个值的条件下 X 的分布。如设 X 表示人的体重，Y 表示人的身高，则 X 与 Y 之间一般有相依关系。现在若限定 $Y=1.7$ 米，在此条件下体重 X 的分布显然与无此限制下 X 的分布是不同的。

1. 离散型随机变量的条件分布列

设二维离散型随机向量(X, Y)的联合分布列为

$$p_{ij}=P(X=x_i, Y=y_j), \ i=1, 2, \cdots, j=1, 2, \cdots$$

仿照条件概率的定义，容易给出如下概念。

定义 3.3.1　对一切使 $P(Y=y_j)=p_{\cdot j}=\sum_i p_{ij}>0$ 的 y_j，称

$$p_{i|j}=P(X=x_i|Y=y_j)=\frac{P(X=x_i, Y=y_j)}{P(Y=y_j)}=\frac{p_{ij}}{p_j}, \ i=1, 2, \cdots$$

为给定 $Y=y_j$ 条件下 X 的分布列。

同样可以定义给定 $X=x_i$ 条件下 Y 的分布列。

例 3.3.1　设二维离散型随机向量(X, Y)的联合分布列为

$X \backslash Y$	1	2	3	$p_{i\cdot}$
1	0.1	0.3	0.2	0.6
2	0.2	0.05	0.15	0.4
$p_{\cdot j}$	0.3	0.35	0.35	1

试求给定 $X=1$ 条件下 Y 的分布列。

解　因为 $P(X=1)=p_{1\cdot}=0.6$，所以用第一行各元素分别除以 0.6，就可以得到给定 $X=1$ 条件下 Y 的条件分布列为

$Y\|X=1$	1	2	3
P	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$

同理可以得到：给定 $X=2$ 条件下 Y 的条件分布列，给定 $Y=1$ 条件下 X 的条

件分布列，给定 $Y=2$ 条件下 X 的条件分布列，给定 $Y=3$ 条件下 X 的条件分布列。

从这个例子看出，二维联合分布列只有一个，而条件分布列有 5 个。若 X 与 Y 的取值更多，则条件分布列也更多。每个条件分布都从一个侧面描述了一种状态下的特定分布。可见条件分布的内容丰富，其应用也更广。

2. 连续型随机变量的条件分布密度函数

设二维连续型随机向量 (X, Y) 的联合密度函数为 $f(x, y)$，边缘密度函数分别为 $f_X(x)$，$f_Y(y)$。

定义 3.3.2 对一切使 $f_Y(y)>0$ 的 y，给定 $Y=y$ 条件下 X 的条件密度函数为

$$f(x|y)=\frac{f(x, y)}{f_Y(y)}$$

对一切使 $f_X(x)>0$ 的 x，给定 $X=x$ 条件下 Y 的条件密度函数为

$$f(y|x)=\frac{f(x, y)}{f_X(x)}$$

例 3.3.2 设 (X, Y) 服从单位圆面上的均匀分布，试求给定 $Y=y$ 条件下 X 的条件密度函数为 $f(x|y)$。

解 因为

$$f(x, y)=\begin{cases}\dfrac{1}{\pi}, & x^2+y^2\leqslant 1\\ 0, & \text{其它}\end{cases}$$

由此得 Y 的边缘密度函数为

$$f_Y(y)=\begin{cases}\dfrac{2}{\pi}\sqrt{1-y^2}, & -1\leqslant y\leqslant 1\\ 0, & \text{其它}\end{cases}$$

所以当 $-1<y<1$ 时，有

$$f(x|y)=\frac{f(x, y)}{f_Y(y)}=\begin{cases}\dfrac{1}{2\sqrt{1-y^2}}, & -\sqrt{1-y^2}\leqslant x\leqslant\sqrt{1-y^2}\\ 0, & \text{其它}\end{cases}$$

若将 $y=0$ 和 $y=0.5$ 分别代入上式可得

$$f(x\mid y=0)=\begin{cases}\dfrac{1}{2}, & -1\leqslant x\leqslant 1\\ 0, & \text{其它}\end{cases}$$

$$f(x \mid y=0.5)=\begin{cases}\dfrac{1}{\sqrt{3}}, & -\dfrac{\sqrt{3}}{2} \leqslant x \leqslant \dfrac{\sqrt{3}}{2} \\ 0, & \text{其它}\end{cases}$$

这是两个均匀分布的密度函数。

进一步可见：当 $-1<y<1$ 时，给定 $Y=y$ 条件下，X 服从 $(-\sqrt{1-y^2}, \sqrt{1-y^2})$ 上的均匀分布。同理有：当 $-1<x<1$ 时，给定 $X=x$ 条件下，Y 服从 $(-\sqrt{1-x^2}, \sqrt{1-x^2})$ 上的均匀分布。

练习 3.3

1. 设二维离散型随机向量 (X, Y) 的联合分布列为

$X \backslash Y$	1	2	3
1	0.1	0.3	0.2
2	0.2	0.1	0.1

试求给定 $X=2$ 条件下 Y 的分布列。

2. 设二维连续型随机向量 (X, Y) 的联合密度函数为

$$f(x, y)=\begin{cases}e^{-x}, & 0<y<x \\ 0, & \text{其它}\end{cases}$$

求条件密度函数 $f(y \mid x)$。

3. 设二维连续型随机向量 (X, Y) 的联合密度函数为

$$f(x, y)=\begin{cases}1, & |y|<x,\ 0<x<1 \\ 0, & \text{其它}\end{cases}$$

求条件密度函数 $f(x \mid y)$。

4. 设随机变量 X 与 Y 相互独立，且 $X \sim (\lambda_1)$，$Y \sim P(\lambda_2)$。试证明在已知 $X+Y=n$ 的条件下，X 服从二项分布 $B(n, p)$，其中 $p=\dfrac{\lambda_1}{\lambda_1+\lambda_2}$。

5. 设在一段时间内进入某一商店的顾客人数 X 服从泊松分布 $P(\lambda)$，每个顾客购买某种物品的概率为 p，并且各个顾客是否购买该种物品相互独立，试证明进入商店的顾客购买这种物品的人数 Y 服从参数为 λp 的泊松分布。

6. 试证明二维正态分布的条件分布是一维正态分布。

第4章　随机变量的数字特征

随机变量的取值是不确定的,但随机变量的变化又有自己的统计规律性。规律性之一就是在前两章介绍的可以具体、细致地描述随机变量的概率分布或概率密度(一般情况下它们很难求得);规律性之二就是在主要概率特征上表现出的稳定性。这种稳定的随机变量的特征往往比全面的概率分布(或概率密度)更直接、更简洁、更清晰、更实用地反映出随机变量的本质。

商店出售的每瓶食油的净含量是不同的,是个随机变量。而每瓶食油的商标上只是标有:净含量5000毫升±5%。这里的5000毫升并不是该瓶食油的实际净含量(逐个标出每瓶油的实际净含量也是不实际的),而是该批食油每瓶的平均净含量(数字特征),±5%是每瓶食油实际净含量与平均净含量的差。实际上,顾客知道了商标上的数字特征就可以对所买的食油有整体的了解。

商店里出售的衣服一般只有几种规格:L,XL,XXL,…,而这些代码和数字正是某类人群身体的尺寸(随机变量)的特征,人们可以根据自己的特征去买衣服。量体裁衣定做衣服当然好,但是没有特殊需要时,大多数人们还是到商店去买。

可见在日常生活中遇到大量的是与随机变量数字特征有关的实际问题。

另外,在后面的学习过程中,我们还会发现随机变量的数字特征大多是概率分布(或概率密度)中的参数,这样一旦知道了数字特征,也就知道了其对应的概率分布(或概率密度)。正因为如此,在数理统计中我们还要对这些数字特征进行估计和检验。

本章主要讨论随机变量的期望、方差、协方差、相关系数及其应用。最后还介绍了作为概率论经典结论和数理统计理论基础的大数定律和中心极限定理。

通过本章的学习,要注意在遇到复杂问题时善于抓特征,看本质,去粗取精,培养一种良好的思维方法和工作方法。

4.1　随机变量的数字特征

1. 数学期望

"平均数"是我们在日常生活中使用最多的一个数字特征，像"平均身高"、"平均亩产量"、"平均产值"、"平均成绩"等等。它简明地指出所研究对象的位置特征，对评判事物，作出决策都有很重要的作用。例如 10 月 1 日某商场准备搞促销活动，统计资料表明，如果在商场内搞促销活动，可获得经济效益 3 万元，在商场外搞促销活动，如果不遇到雨天可获经济效益 12 万元，如果遇到雨天则带来经济损失 5 万元。9 月 30 日的天气预报称当地有雨的概率为 40%，那么商场应该选择哪种促销方式呢？显然在商场外搞促销活动的经济效益 X 是个随机变量，其概率分布为

$$P(X=12)=0.6,\quad P(X=-5)=0.4$$

要作出决策就要将此时的平均效益与 3 万元比较。又如何求平均效益呢？要想全面客观地反映平均效益就须既要考虑 X 的所有取值，又要考虑 X 取每一个值时的概率，即为

$$12\times0.6+(-5)\times0.4=5.2(\text{万元})$$

平均效益 5.2 万元即为 X 的数学期望。

(1) 离散型随机变量的数学期望

定义 4.1.1　设离散型随机变量 X 的概率分布为

$$P(X=x_k)=p_k\qquad(k=1,\ 2,\ \cdots)$$

如果级数 $\sum_{k=1}^{\infty}x_kp_k$ 绝对收敛，则称 $\sum_{k=1}^{\infty}x_kp_k$ 为 X 的**数学期望(mathematical expectation)**，简称**期望或均值(mean)**，记作 $E(X)$，即 $E(X)=\sum_{k=1}^{\infty}x_kp_k$。如果 $\sum_{k=1}^{\infty}x_kp_k$ 不是绝对收敛，即级数 $\sum_{k=1}^{\infty}|x_k|p_k$ 发散，则称 X 的数学期望不存在。

定义中"绝对收敛"这一条件，是为了保证不管 x_k 的次序如何改变，级数都应收敛，且收敛于同一数值。显然对于一个随机变量的分布列来说，这一要求是合理的。

数学期望实际上是以概率 p_k 为权的加权平均值，当 p_k 为同一数值时，数学期望即为算术平均值。数学期望是不变的常数，而不再随机。

例 4.1.1　设随机变量 X 服从参数为 p 的 0－1 分布，求 $E(X)$。

解 由题意知，X 的分布列为

$$P(X=k)=p^k(1-p)^{1-k} \quad (k=0,\ 1)$$

于是 $E(X)=0\cdot(1-p)+1\cdot p=p$。

例 4.1.2 设随机变量 X 服从二项分布，即 $X\sim B(n,\ p)$，求 $E(X)$。

解 由上题知 0-1 分布的期望是 p，而二项分布是 n 个相互独立的具有相同概率 p 的 0-1 分布的和。可以猜想一下，二项分布的数学期望是否为 np？下面我们用定义来计算。

依题知 X 的分布列为

$$P(X=k)=\mathrm{C}_n^k p^k(1-p)^{n-k},\quad k=0,\ 1,\ 2,\ \cdots,n$$

于是有

$$\begin{aligned}E(X)&=\sum_{k=0}^{n}k\cdot \mathrm{C}_n^k p^k(1-p)^{n-k}\\&=\sum_{k=1}^{n}k\cdot\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}\\&=\sum_{k=1}^{n}\frac{np(n-1)!}{(k-1)![(n-1)-(k-1)]!}p^{k-1}(1-p)^{(n-1)-(k-1)}\\&\xlongequal{\text{令 } k'=k-1} np\sum_{k'=0}^{n-1}\mathrm{C}_{n-1}^{k'}p^{k'}(1-p)^{(n-1)-k'}\\&=np\end{aligned}$$

最后一个等式成立可以从下面两方面来证明。

一方面，由二项式展开公式有：

$$\sum_{k'=0}^{n-1}\mathrm{C}_{n-1}^{k'}p^{k'}(1-p)^{(n-1)-k'}=[p+(1-p)]^{n-1}=1$$

另一方面，$\sum_{k'=0}^{n-1}\mathrm{C}_{n-1}^{k'}p^{k'}(1-p)^{(n-1)-k'}$ 可以看成是二项分布 $B(n-1,\ p)$ 中所有概率之和。由概率分布的性质，其值等于 1。利用这一性质计算随机变量的数字特征是常用的方法。

例 4.1.3 设随机变量 X 服从 Poisson 分布，即 $X\sim P(\lambda)$，求 $E(X)$。

解 求解的过程和道理与上题类似，此处不再详加说明。

依题知 X 的分布列为

$$P(X=k)=\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}\quad(\lambda>0,\ k=0,\ 1,\ 2,\ \cdots)$$

于是有

$$E(X)=\sum_{k=0}^{\infty}k\cdot\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}=\lambda\mathrm{e}^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}$$

$$=\lambda e^{-\lambda}\sum_{k'=0}^{\infty}\frac{\lambda^{k'}}{k'!}=\lambda e^{-\lambda}\cdot e^{\lambda}=\lambda$$

例 4.1.4　某种产品的每件表面上的疵点数服从 Poisson 分布，平均每件上有 0.8 个疵点。若规定疵点数不超过 1 个为一等品，价值 10 元；疵点数大于 1 个不多于 4 个为二等品，价值 8 元；疵点数超过 4 个为废品；求：①产品的废品率；②产品价值的平均值。

解　设 X 代表每件产品上的疵点数，由题意知：$E(X)=0.8$，即 $\lambda=0.8$。

① 因为 $P(X>4)=1-P(X\leqslant 4)=1-\sum_{k=0}^{4}\frac{0.8^k}{k!}e^{-0.8}=0.001412$，所以产品的废品率为 0.001412。

②设 Y 代表产品的价值，那么 Y 的概率分布为

Y	10	8	0
P	$P(X\leqslant 1)$	$P(1<X\leqslant 4)$	$P(X>4)$

所以产品价值的平均值为

$$E(Y)=10\times P(X\leqslant 1)+8\times P(1<X\leqslant 4)+0\times P(X>4)$$

$$=10\times\sum_{k=0}^{1}\frac{0.8^k}{k!}e^{-0.8}+8\times\sum_{k=2}^{4}\frac{0.8^k}{k!}e^{-0.8}+0=9.61(\text{元})$$

(2)连续型随机变量的数学期望

定义 4.1.2　设连续型随机变量 X 的概率密度为 $f(x)$，如果广义积分 $\int_{-\infty}^{+\infty}xf(x)dx$ 绝对收敛，则称 $\int_{-\infty}^{+\infty}xf(x)dx$ 为 X 的数学期望，简称期望或均值，记作 $E(X)$，即 $E(X)=\int_{-\infty}^{+\infty}xf(x)dx$。

连续型随机变量数学期望的定义是离散型随机变量数学期望定义的拓展，只要应用定积分的定义和中值定理就可以得到。它只不过是将离散的级数运算变为相应的连续的积分运算。

例 4.1.5　设随机变量 X 服从$[a,b]$上的均匀分布，求 $E(X)$。

解　由题意知，X 的概率密度函数为

$$f(x)=\begin{cases}\frac{1}{b-a}, & a\leqslant x\leqslant b\\ 0, & \text{其它}\end{cases}$$

于是有

$$E(X)=\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x=\int_a^b\frac{x}{b-a}\mathrm{d}x=\frac{a+b}{2}$$

例 4.1.6 设随机变量 X 服从参数为 λ 的指数分布，求 $E(X)$。

解 由题意知，X 的概率密度函数为

$$f(x)=\begin{cases}\lambda\mathrm{e}^{-\lambda x}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

于是有

$$\begin{aligned}E(X)&=\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x=\int_0^{+\infty}\lambda x\mathrm{e}^{-\lambda x}\mathrm{d}x\\&=-x\mathrm{e}^{-\lambda x}\Big|_0^{+\infty}+\int_0^{+\infty}\mathrm{e}^{-\lambda x}\mathrm{d}x\\&=\frac{1}{\lambda}\int_0^{+\infty}\lambda\mathrm{e}^{-\lambda x}\mathrm{d}x\\&=\frac{1}{\lambda}\end{aligned}$$

最后一个等式又用到概率密度的性质 $\int_{-\infty}^{+\infty}f(x)\mathrm{d}x=1$。

例 4.1.7 设随机变量 X 服从正态分布 $N(\mu,\sigma^2)$，求 $E(X)$。

解 由题意知，X 的概率密度函数为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}},\quad -\infty<x<+\infty$$

于是有

$$\begin{aligned}E(X)&=\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x=\int_{-\infty}^{+\infty}x\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x\\&\xlongequal{y=\frac{x-\mu}{\sigma}}\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty}y\mathrm{e}^{-\frac{y^2}{2}}\mathrm{d}y+\mu\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{y^2}{2}}\mathrm{d}y\\&=0+\mu\\&=\mu\end{aligned}$$

特别地，当 $X\sim N(0,1)$ 时，$E(X)=0$。

以上推导的几种常见重要分布的数学期望都和它们分布中的参数有关，以正态分布为例，参数 μ 正是它的期望。我们知道，正态分布的密度函数曲线以直线 $x=\mu$ 为对称轴，当 $x=\mu$ 时，$f(x)$ 达到最大值。因此这一结果从直观上说，也是很自然的(如图 4.1.1)。

例 4.1.8 设随机变量 $X\sim f(x)$，$E(X)=\frac{7}{12}$，且

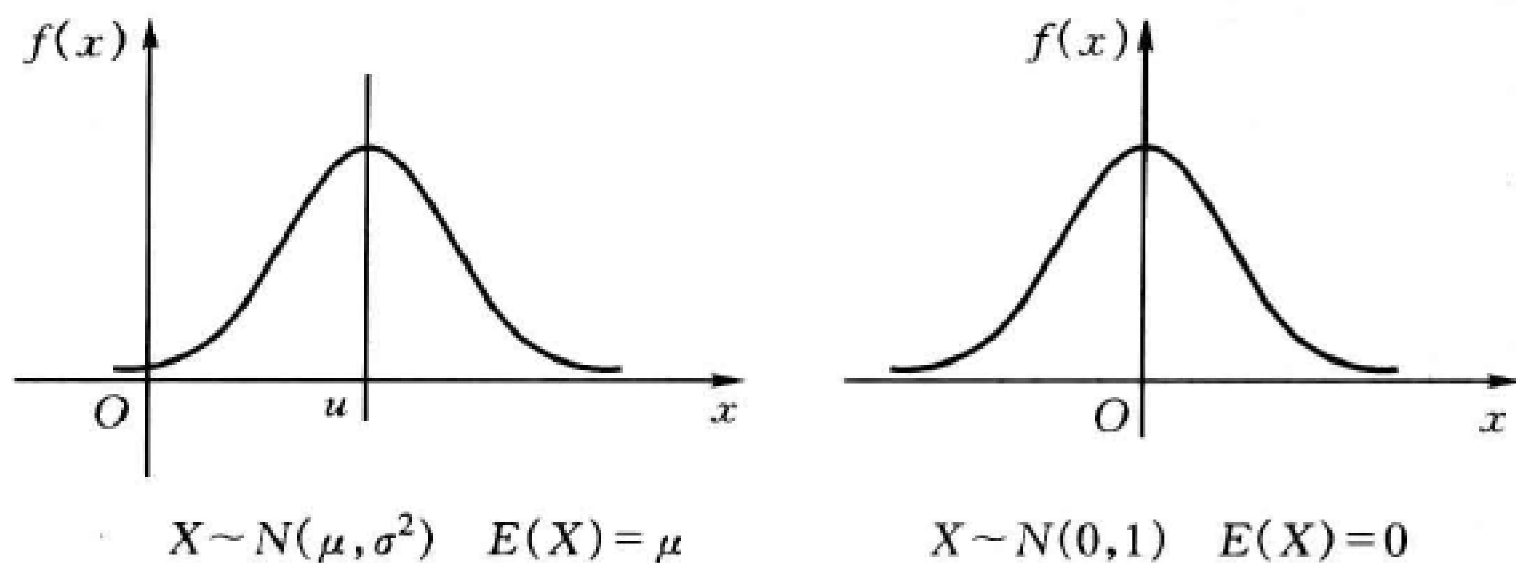


图 4.1.1

$$f(x)=\begin{cases} ax+b, & 0\leqslant x\leqslant 1 \\ 0, & \text{其它} \end{cases}$$

求 a 与 b 的值；并求分布函数 $F(x)$。

解　由题意知

$$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x=\int_0^1 (ax+b)\mathrm{d}x=\frac{a}{2}+b=1$$

$$E(X)=\int_{-\infty}^{+\infty} xf(x)\mathrm{d}x=\int_0^1 x(ax+b)\mathrm{d}x=\frac{a}{3}+\frac{b}{2}=\frac{7}{12}$$

解方程组得 $a=1, b=\frac{1}{2}$。

当 $0\leqslant x<1$ 时，有

$$F(x)=\int_{-\infty}^{x} f(t)\mathrm{d}t=\int_0^x \left(t+\frac{1}{2}\right)\mathrm{d}t=\frac{x^2}{2}+\frac{x}{2}$$

所以

$$F(x)=\begin{cases} 0, & x<0 \\ \frac{1}{2}(x^2+x), & 0\leqslant x<1 \\ 1, & x\geqslant 1 \end{cases}$$

应该指出，并非所有随机变量都有数学期望，例如 X 的密度函数为

$$f(x)=\frac{1}{\pi(1+x^2)},\quad -\infty<x<+\infty$$

由于广义积分

$$\int_{-\infty}^{+\infty} |x| f(x)\mathrm{d}x=\frac{2}{\pi}\int_0^{+\infty}\frac{x}{1+x^2}\mathrm{d}x=\frac{1}{\pi}\ln(1+x^2)\Big|_0^{+\infty}=+\infty$$

发散，所以 $E(X)$ 不存在。

(3)随机变量函数的数学期望

在实际问题中，我们经常需要计算随机变量函数的数学期望。下面不加证明给出几个有关的定理。

定理 4.1.1 设 X 是随机变量，$Y=g(X)$，并且 $E(Y)$存在，则

① 若 X 为离散型随机变量，其概率分布为

$$P(X=x_k)=p_k,\quad k=1,\ 2,\ \cdots$$

则 Y 的数学期望为

$$E(Y)=E(g(X))=\sum_{k=1}^{\infty}g(x_k)p_k$$

② 若 X 为连续型随机变量，其概率密度为 $f(x)$，则 Y 的数学期望为

$$E(Y)=E(g(X))=\int_{-\infty}^{+\infty}g(x)f(x)\mathrm{d}x$$

这两个公式的重要性在于：当计算 $g(X)$的期望时，不必先求出 $g(X)$的分布，而直接利用 X 的分布来计算 $E[g(X)]$，很方便。对于随机向量来讲，也有类似的结论。

定理 4.1.2 设$(X,\ Y)$是二维随机向量，Z 是 X，Y 的函数 $Z=g(X,\ Y)$，并且 $E(Z)$存在，则

① 若$(X,\ Y)$为离散型随机向量，其概率分布为

$$P(X=x_i,\ Y=y_j)=p_{ij},\ i,\ j=1,\ 2,\ \cdots$$

则 Z 的数学期望为

$$E(Z)=E(g(X,\ Y))=\sum_{j=1}^{\infty}\sum_{i=1}^{\infty}g(x_i,\ y_j)p_{ij}$$

② 若$(X,\ Y)$为连续型随机向量，其联合密度函数为 $f(x,\ y)$，则 Z 的数学期望为

$$E(Z)=E(g(X,\ Y))=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}g(x,\ y)f(x,\ y)\mathrm{d}x\mathrm{d}y$$

例 4.1.9 设随机变量 X 在$[0,\ \pi]$上服从均匀分布，求 $E(\sin X)$，$E(X^2)$及 $E[X-E(X)]^2$。

解 由定理 4.1.1，有

$$E(\sin X)=\int_{-\infty}^{+\infty}\sin x f(x)\mathrm{d}x=\int_0^{\pi}\sin x\cdot\frac{1}{\pi}\mathrm{d}x=\frac{1}{\pi}(-\cos x)\Big|_0^{\pi}=\frac{2}{\pi}$$

$$E(X^2)=\int_{-\infty}^{+\infty}x^2f(x)\mathrm{d}x=\int_0^{\pi}x^2\cdot\frac{1}{\pi}\mathrm{d}x=\frac{\pi^2}{3}$$

$$E[X-E(X)]^2=E[X-\frac{\pi}{2}]^2=\int_0^{\pi}(x-\frac{\pi}{2})^2\cdot\frac{1}{\pi}\mathrm{d}x=\frac{\pi^2}{12}$$

对于上面计算 $E(X^2)$及 $E[(X-E)]^2$ 的方法，在后面方差的计算中还要用到，请大家予以重视。

例 4.1.10 假定国际市场每年对我国某种商品的需求量是一个随机变量 X

（单位：吨），它服从[2000，4000]上均匀分布，已知该商品每售出 1 吨，可获 3 万元的外汇，但若销售不出去，则每吨要损失各种费用 1 万元，那么如何组织货源，才可使收益最大？

解　设 y 为组织的货源数量，Y 为收益，显然收益 Y 是销售量 X 和组织货源数量 y 的函数，由于 X 是随机变量，所以收益 Y 也是一个随机变量（组织货源量 y 不是随机变量）。由于 X 服从[2000，4000]上的均匀分布，因此只需在[2000，4000]上考虑。依题有

$$Y=g(X)=\begin{cases} 3y, & \text{当 } X\geqslant y \text{ 时} \\ 3X-(y-X), & \text{当 } X<y \text{ 时} \end{cases}$$

于是

$$\begin{aligned} E(Y)=E(g(X))&=\int_{-\infty}^{+\infty}g(x)f(x)\mathrm{d}x=\frac{1}{2000}\int_{2000}^{4000}g(x)\mathrm{d}x \\ &=\frac{1}{2000}\int_{2000}^{y}(4x-y)\mathrm{d}x+\frac{1}{2000}\int_{y}^{4000}3y\mathrm{d}x \\ &=\frac{1}{1000}(-y^2+7000y-4000000) \end{aligned}$$

由微积分知识可算出：当 $y=3500$ 时 $E(Y)$ 最大，因此应组织 3500 吨的商品。

例 4.1.11　设某商店经营一种商品，每周的进货量 X 和顾客对该种商品的需求量 Y 是两个相互独立的随机变量，均服从[10，20]上均匀分布，此商店每售出一个单位的商品可获利 1000 元，若需求量超过进货量，可从其它商店调剂供应，此时售出的每单位商品仅获利 500 元，求此商店经销这种商品每周获利的期望。

解　此题与上例不同，这是一个二维随机向量问题。设此商店经销该商品每周可获利 L 元。依题有

$$L=\begin{cases} 1000Y, & Y\leqslant X \\ 1000X+500(Y-X), & Y>X \end{cases}$$

而(X, Y)的联合概率密度为

$$f(x, y)=\begin{cases} \dfrac{1}{100}, & 10\leqslant x\leqslant 20,\ 10\leqslant y\leqslant 20 \\ 0, & \text{其它} \end{cases}$$

所以

$$\begin{aligned} E(L)&=\iint_D L\cdot f(x, y)\mathrm{d}x\mathrm{d}y \\ &=\iint_{D_1}1000y\times\frac{1}{100}\mathrm{d}x\mathrm{d}y+\iint_{D_2}500(x+y)\times\frac{1}{100}\mathrm{d}x\mathrm{d}y \end{aligned}$$

$$= \int_0^{20} dx \int_0^x 10y dy + \int_0^{20} dx \int_x^{20} 5(x+y) dy$$

$$= 5\int_0^{20} (x^2 - 100) dx + \frac{5}{2}\int_0^{20} [(x+20)^2 - 4x^2] dx$$

$$= \frac{20000}{3} + 7500$$

$$= 14167$$

故该商店经销这种商品每周的期望获利 14167 元。

例 4.1.12 设(X,Y)的联合概率分布为：

Y \ X	0	1	2	3
1	0	$\frac{3}{8}$	$\frac{3}{8}$	0
3	$\frac{1}{8}$	0	0	$\frac{1}{8}$

求 $E(X)$，$E(Y)$，$E(X \cdot Y)$。

解 要求 $E(X)$和 $E(Y)$，需先求出 X 和 Y 的边缘分布。关于 X 和 Y 的边缘分布为

X	1	3
P	$\frac{3}{4}$	$\frac{1}{4}$

Y	0	1	2	3
P	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

那么有

$$E(X) = 1 \times \frac{3}{4} + 3 \times \frac{1}{4} = \frac{3}{2};$$

$$E(Y) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{3}{2};$$

$$E(X \cdot Y) = (1 \times 0) \times 0 + (1 \times 1) \times \frac{3}{8} + (1 \times 2) \times \frac{3}{8} + (1 \times 3) \times 0 + (3 \times 0) \times \frac{1}{8} + (3 \times 1) \times 0 + (3 \times 2) \times 0 + (3 \times 3) \times \frac{1}{8}$$

$$= \frac{9}{4}$$

对于上面计算 $E(X \cdot Y)$的方法，后面协方差的计算中还要用到，请读者予以

重视。

(4)数学期望的性质

从以上有关数学期望的定义和定理出发,我们进一步来讨论数学期望的一些其它性质。假设下面所用到的期望都存在。

性质 1　如果 C 是一个常数,则

$$E(C)=C$$

性质 2　如果 X 是随机变量,a,b 是常数,则

$$E(aX+b)=aE(X)+b$$

性质 3　如果(X,Y)是二维随机向量,则

$$E(X\pm Y)=E(X)\pm E(Y)$$

性质 4　如果(X, Y)是二维随机向量,且 X 和 Y 相互独立,则

$$E(X\cdot Y)=E(X)\cdot E(Y)$$

证　这里只证明性质 2 和性质 4,性质 1 和性质 3 留给读者。

性质 2 的证明(这里只对离散型的情形证明)。

设 X 的概率分布为 $P(X=x_k)=p_k(k=1, 2, \cdots)$,由定理 4.1.1 有

$$\begin{aligned}E(aX+b)&=\sum_{k=1}^{\infty}(ax_k+b)p_k=a\sum_{k=1}^{\infty}x_kp_k+b\sum_{k=1}^{\infty}p_k\\&=aE(X)+b\end{aligned}$$

性质 4 的证明(这里只对连续型的情形证明)。

设(X, Y)的联合密度函数为 $f(x,y)$,其边缘概率密度分别 $f_X(x)$和 $f_Y(y)$,由定理 4.1.2 知

$$E(X\cdot Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xyf(x,y)\,\mathrm{d}x\mathrm{d}y$$

因为 X 和 Y 相互独立,$f(x,y)=f_X(x)\cdot f_Y(y)$,所以有

$$\begin{aligned}E(X\cdot Y)&=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xyf_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y\\&=\int_{-\infty}^{+\infty}xf_X(x)\mathrm{d}x\cdot\int_{-\infty}^{+\infty}yf_Y(y)\mathrm{d}y\\&=E(X)\cdot E(Y)\end{aligned}$$

对于性质 3 和性质 4 可以推广到有限个随机变量的情形,即对于 $n>2$,同样有

$$E(X_1+X_2+\cdots X_n)=E(X_1)+E(X_2)+\cdots+E(X_n)$$

而当 X_1, X_2, $\cdots$, X_n 相互独立时,类似地有

$$E(X_1\cdot X_2\cdot\cdots\cdot X_n)=E(X_1)\cdot E(X_2)\cdot\cdots\cdot E(X_n)$$

性质 4 的逆命题不成立，即：如果 X 与 Y 满足 $E(X\cdot Y)=E(X)\cdot E(Y)$，那么 X 与 Y 不一定相互独立。例如在例 4.1.12 中 $E(X\cdot Y)=E(X)\cdot E(Y)=\frac{9}{4}$，但 $P(X=1,\ Y=0)=0$，而 $P(X=1)\cdot P(Y=0)=\frac{3}{32}$，所以 X 与 Y 不独立。

例 4.1.13　试用性质计算例 4.1.9 中的 $E[X-E(X)]^2$。

解　由性质知

$$\begin{aligned}E[X-E(X)]^2&=E[X^2-2E(X)\cdot X+E^2(X)]\\&=E(X^2)-2E(X)\cdot E(X)+E^2(X)\\&=E(X^2)-E^2(X)\\&=\frac{\pi^2}{3}-\left(\frac{\pi}{2}\right)^2\\&=\frac{\pi^2}{12}\end{aligned}$$

例 4.1.14　一次数学测验由 40 个单项选择题构成，每个选择题有 4 个选项，每题选择正确答案得 2.5 分，否则 0 分，满分 100 分。学生甲选对任一题的概率为 0.8，学生乙则每次都从 4 个选项中任选一个。分别求学生甲和乙在这次数学测验中的期望成绩。

解　设学生甲、乙在测验中选对题的个数分别为 X 与 Y，则所得的成绩分别为 $2.5X$ 与 $2.5Y$。由于 $X\sim B(40,\ 0.8)$，$Y\sim B(40,\ 0.25)$，所以

$$E(X)=40\times 0.8=32,\ E(Y)=40\times 0.25=10$$

于是学生甲和乙的期望成绩分别为

$$E(2.5X)=2.5\times E(X)=2.5\times 32=80$$
$$E(2.5Y)=2.5\times E(Y)=2.5\times 10=25$$

2. 方　差

随机变量取值的稳定性是判断随机现象性质的十分重要的指标。例如某地区地震仪上描出的曲线如果起伏很大，这说明该地区地下活动异常，是地震的预兆；某天股市中股票价格出现异常波动，这就预示着社会经济中将有重大事件发生；一台仪器在测量某一元件的某数量指标时，若在多次测量中数据的差异很大，则说明该仪器存在质量问题，需修理或更新了。因此如何衡量随机变量的稳定特征在现实生活中意义重大。

(1) 方差的基本概念和性质

定义 4.1.3　设 X 是随机变量，期望 $E(X)$ 存在，如果 $E[X-E(X)]^2$ 存在，则

$E[X-E(X)]^2$ 称为 X 的**方差(variance)**，记作 $D(X)$，即

$$D(X)=E[X-E(X)]^2$$

而 $\sqrt{D(X)}$ 称为 X 的**标准差(standard deviation)**。

由方差定义的数学表达式可以看出，方差实际上是随机变量 X 与它均值 $E(X)$ 差的平方的期望值，它的大小自然可以衡量随机变量的稳定状态，所以方差反映了随机变量的变异特征。对于一个随机变量来讲，方差 $D(X)$ 是一个稳定常数，不再是随机的了。

如果 X 是离散型随机变量，且其概率分布为

$$P(X=x_k)=p_k(k=1, 2, \cdots)$$

那么

$$D(X) = \sum_{k=1}^{\infty}[x_k - E(X)]^2 p_k$$

如果 X 是连续型随机变量，其概率密度为 $f(x)$，那么

$$D(X) = \int_{-\infty}^{+\infty}[x - E(X)]^2 f(x) Dx$$

在前面的例 4.1.13 中，实际上我们已经证明了一个很重要的计算方差的公式

$$D(X) = E(X^2) - E^2(X)$$

在后面有关方差的计算和证明时，经常要用到上面这个公式。

由方差的定义，我们可以得出关于方差的一些性质(假设下面所用到的方差都存在)。

性质 1　如果 C 是一个常数，则

$$D(C)=0$$

性质 2　如果 C 是一个常数，则

$$D(X+C)=D(X)$$

性质 3　如果 a 是一个常数，则

$$D(aX)=a^2D(X)$$

性质 4　设 X 和 Y 相互独立，则

$$D(X\pm Y)=D(X)+D(Y)$$

证　(性质 1)

$$D(C)=E[C-E(C)]^2=E(C-C)^2=0$$

(性质 2)

$$D(X+C)=E[(X+C)-E(X+C)]^2=E[X-E(X)]^2=D(X)$$

下面我们用公式 $D(X)=E(X^2)-E^2(X)$ 来证明性质 3 和性质 4。

(性质 3)

$$D(aX)=E[(aX)^2]-E^2(aX)=a^2[E(X^2)-E^2(X)]=a^2D(X)$$

（性质 4）

$$\begin{aligned}D(X\pm Y)&=E(X\pm Y)^2-E^2(X\pm Y)\\&=E(X^2\pm 2XY+Y^2)-[E^2(X)\pm 2E(X)E(Y)+E^2(Y)]\\&=E(X)^2-E^2(X)+E(Y)^2-E^2(Y)\pm 2[E(XY)-E(X)E(Y)]\end{aligned}$$

由于 X 和 Y 相互独立，有 $E(XY)=E(X)E(Y)$，所以

$$D(X\pm Y)=E(X)^2-E^2(X)+E(Y)^2-E^2(Y)=D(X)+D(Y)$$

一般情形下，有

$$D(X\pm Y)=D(X)+D(Y)\pm 2[E(XY)-E(X)E(Y)]$$

性质 4 的结论可进一步推广如下（证明略）。

如果 $X_1, X_2, \cdots, X_n$ 相互独立，$c_i(i=1, 2, \cdots, n)$是任意常数，那么

$$D\left(\sum_{i=1}^{n}c_iX_i\right)=\sum_{i=1}^{n}c_i^2D(X_i)$$

(2)方差的计算

例 4.1.15 设随机变量 X 服从参数为 p 的 0-1 分布，求 $D(X)$。

解 由题意知，X 的概率分布为

$$P(X=k)=p^k(1-p)^{1-k}\quad(k=0, 1)$$

且 $E(X)=p$，又因为

$$E(X^2)=0^2\cdot(1-p)+1^2\cdot p=p$$

所以 $D(X)=E(X^2)-E^2(X)=p-p^2=p(1-p)$。

例 4.1.16 设随机变量 X 服从参数为λ 的 Poisson 分布，求 $D(X)$。

解 已知 X 的概率分布为

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}(\lambda>0, k=0, 1, 2,\cdots)$$

且 $E(X)=\lambda$，又因为

$$\begin{aligned}E(X^2)&=\sum_{k=0}^{\infty}k^2\cdot\frac{\lambda^k}{k!}e^{-\lambda}=\sum_{k=0}^{\infty}[k(k-1)+k]\cdot\frac{\lambda^k}{k!}e^{-\lambda}\\&=\sum_{k=0}^{\infty}k(k-1)\frac{\lambda^k}{k!}e^{-\lambda}+\sum_{k=0}^{\infty}k\frac{\lambda^k}{k!}e^{-\lambda}\\&=\lambda^2\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}e^{-\lambda}+\lambda\\&\xlongequal{k'=k-2}\lambda^2\sum_{k'=0}^{\infty}\frac{\lambda^{k'}}{k'!}e^{-\lambda}+\lambda\end{aligned}$$

$$= \lambda^2 + \lambda$$

所以 $D(X)=E(X^2)-E^2(X)=\lambda^2+\lambda-\lambda^2=\lambda$。

例 4.1.17　设随机变量 X 服从参数为 n，p 的二项分布，求 $D(X)$。

解　在这里暂时不用方差的定义求 $D(X)$（留给读者自己练习），而从另一角度考虑。

我们知道，如果 X 服从二项分布，则 X 可以写成 $X=\sum_{i=1}^{n} X_i$，其中 X_i 相互独立，且 $P(X_i=1)=p$，$P(X_i=0)=1-p$ $(i=1, 2,\cdots,)$，那么

$$D(X)=D\left(\sum_{i=1}^{n} X_i\right)=\sum_{i=1}^{n} D(X_i)=np(1-p)$$

例 4.1.18　设随机变量 X 服从$[a, b]$上的均匀分布，求 $D(X)$。

解　已知 X 的概率密度函数为

$$f(x)=\begin{cases} \dfrac{1}{b-a}, & a\leqslant x\leqslant b \\ 0, & \text{其它} \end{cases}$$

而且 $E(X)=\dfrac{a+b}{2}$，又因为

$$E(X^2)=\int_{-\infty}^{+\infty} x^2 f(x)\mathrm{d}x=\int_a^b \frac{x^2}{b-a}\mathrm{d}x$$

$$=\frac{1}{3}(a^2+ab+b^2)$$

所以

$$D(X)=\frac{1}{3}(a^2+ab+b^2)-\left(\frac{a+b}{2}\right)^2=\frac{(b-a)^2}{12}$$

例 4.1.19　设随机变量 X 服从参数为 $\lambda(>0)$ 的指数分布，求 $D(X)$。

解　由题意知，X 的概率密度函数为

$$f(x)=\begin{cases} \lambda \mathrm{e}^{-\lambda x}, & x>0 \\ 0, & x\leqslant 0 \end{cases}$$

而 $E(X)=\dfrac{1}{\lambda}$，又因为

$$E(X^2)=\int_{-\infty}^{+\infty} x^2 f(x)\mathrm{d}x=\int_0^{+\infty} \lambda x^2 \mathrm{e}^{-\lambda x}\mathrm{d}x$$

$$=-x^2\mathrm{e}^{-\lambda x}\Big|_0^{+\infty}+\frac{2}{\lambda}\int_0^{+\infty}\lambda x \mathrm{e}^{-\lambda x}\mathrm{d}x$$

$$=\frac{2}{\lambda}\cdot\frac{1}{\lambda}=\frac{2}{\lambda^2}$$

所以

$$D(X)=\frac{2}{\lambda^2}-\left(\frac{1}{\lambda}\right)^2=\frac{1}{\lambda^2}$$

例 4.1.20 设随机变量 X 服从正态分布 $N(\mu,\sigma^2)$，求 $D(X)$。

解 直接采用方差的定义计算 $D(X)$。

已知 X 的概率密度函数为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ -\infty<x<+\infty$$

且 $E(X)=\mu$，所以

$$D(X)=E[X-E(X)]^2=\int_{-\infty}^{+\infty}(x-\mu)^2\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

$$\xlongequal{y=\frac{x-\mu}{\sigma}}\int_{-\infty}^{+\infty}y^2\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy$$

$$=\frac{\sigma^2}{\sqrt{2\pi}}\left[y^2e^{-\frac{y^2}{2}}\Big|_{-\infty}^{+\infty}+\int_{-\infty}^{+\infty}e^{-\frac{y^2}{2}}dy\right]$$

$$=\sigma^2\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy=\sigma^2$$

以上几个例题中求到的一些常见分布的方差都与它们分布中的参数有关，一旦求出了期望和方差，它们的分布也就唯一确定了。方差概率意义的直观背景可以通过正态分布中不同的 σ^2 密度曲线清楚地显示出来(图 4.1.2)。

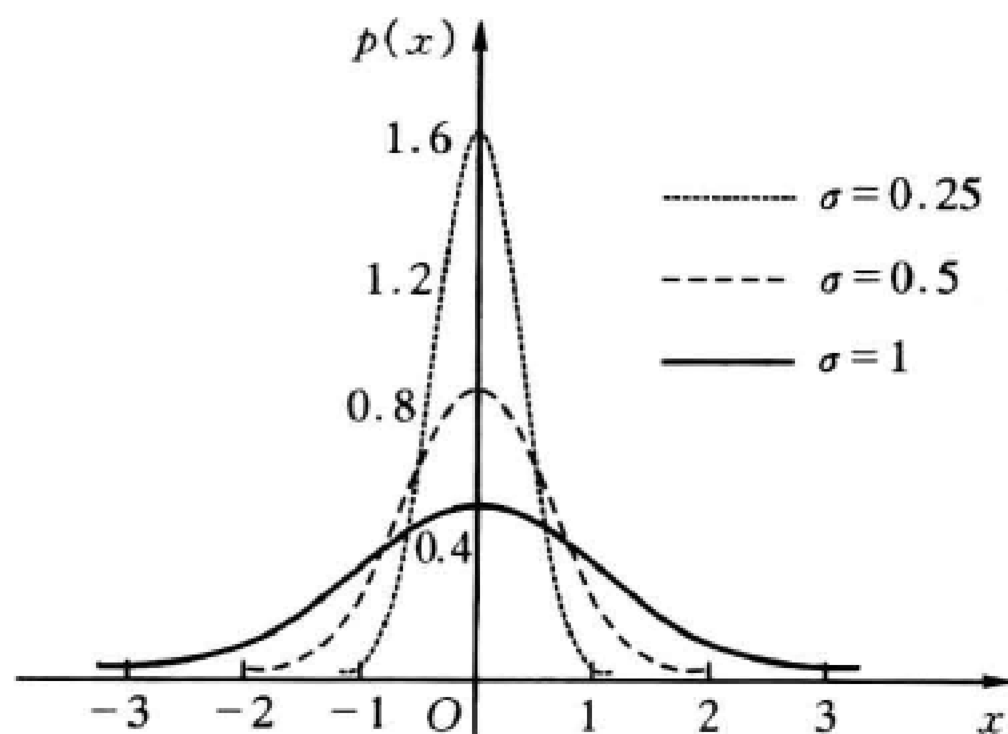


图 4.1.2 σ^2 越小取值越集中

例 4.1.21 设 $f(x)=E(X-x)^2$，$x\in\mathbf{R}$，证明：当 $x=E(X)$ 时，$f(x)$ 达到最小值。

证　依题 $f(x)=E(X-x)^2=E(X^2)-2xE(X)+x^2$，两边对 x 求导数有

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x}=2x-2E(X)$$

显然当 $x=E(X)$ 时，$\frac{\mathrm{d}f(x)}{\mathrm{d}x}=0$。又因 $\frac{\mathrm{d}^2f(x)}{\mathrm{d}x^2}=2>0$，所以当 $x=E(X)$ 时 $f(x)$ 达到最小值，最小值为

$$f(E(X))=E(X-E(X))^2=D(X)$$

这个例子又一次说明了数学期望 $E(X)$ 是随机变量 X 取值的集中位置，反映了 X 的平均值。

例 4.1.22　设随机变量存在数学期望 $E(X)$ 和方差 $D(X)$，试证明：对任意的 $\varepsilon>0$，有

$$P(|X-E(X)|\geqslant\varepsilon)\leqslant\frac{D(X)}{\varepsilon^2}$$

证　（只给出连续型的证明）

$$\begin{aligned}D(X)&=\int_{-\infty}^{+\infty}[x-E(X)]^2f(x)\mathrm{d}x\\&\geqslant\int_{|x-E(X)|\geqslant\varepsilon}[x-E(X)]^2f(x)\mathrm{d}x\\&\geqslant\int_{|x-E(X)|\geqslant\varepsilon}\varepsilon^2f(x)\mathrm{d}x=\varepsilon^2\int_{|x-E(X)|\geqslant\varepsilon}f(x)\mathrm{d}x\\&=\varepsilon^2P[|X-E(X)|\geqslant\varepsilon]\end{aligned}$$

所以

$$P(|X-E(X)|\geqslant\varepsilon)\leqslant\frac{D(X)}{\varepsilon^2}$$

这是著名的**切比雪夫(Chebyshev)**不等式，它表明：方差 $D(X)$ 越小，事件 $(|X-E(X)|\geqslant\varepsilon)$ 发生的概率就越小，即事件 $(|X-E(X)|<\varepsilon)$ 发生的概率越大，X 的取值越集中在 $E(X)$ 的附近。从这个不等式中我们也进一步看出，方差 $D(X)$ 刻画了随机变量 X 的分散程度，是反映 X 波动状态的重要指标。另外切比雪夫不等式也是在第 4.3 节中大数定理的理论基础。

切比雪夫不等式的另一等价形式是：

$$P(|X-E(X)|<\varepsilon)>1-\frac{D(X)}{\varepsilon^2}$$

练习 4.1

1. 一箱产品 20 件，其中 5 件优质品，每次抽取 1 件，共抽取 2 次，求取到的优

质品件数 X 的数学期望(分两种情况讨论:①有放回地抽取;②不放回地抽取)。

2. 试计算上题中 X 的方差。

3. 设 X 的概率分布为

X	4	6	x_3
P	0.5	0.3	a

且 $E(X)=8$,求 x_3 和 a 的值。

4. 据统计,一位 60 岁的健康(一般体检未发生病症)者,在 5 年之内仍然活着和自杀死亡的概率为 $p(0<p<1, p$ 为已知),在 5 年之内非自杀死亡的概率为 $1-p$。保险公司开办 5 年人寿保险,条件是参加者需交纳人寿保险 a 元(a 已知),若 5 年内死亡,公司赔偿 b 元($b>a$),应如何确定才能使公司可期望获益。若有 m 人参加保险,公司可期望从中收益多少?

5. 设连续型随机变量 X 的概率密度为

$$f(x)=\begin{cases} kx^{\alpha}, & 0<x<1 \\ 0, & 其它 \end{cases}$$

其中 $k, \alpha>0$。又已知 $E(X)=0.75$,求 k, α 的值。

6. 设连续型随机变量 X 的概率密度为

$$f(x)=\begin{cases} x, & 0\leqslant x\leqslant 1 \\ 2-x, & 1<x<2 \\ 0, & 其它 \end{cases}$$

试求 $E(X)$ 和 $D(X)$。

7. 设连续型随机变量 X 与 Y 独立,其概率密度分别为

$$f_X(x)=\begin{cases} \sigma e^{-\sigma x}, & x>0 \\ 0, & x\leqslant 0 \end{cases}$$

$$f_Y(y)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{y^2}{2\sigma^2}}, -\infty<y<+\infty$$

其中 $\sigma>0$。记 $Z=2X-3Y+1$,试求 $E(Z)$ 和 $D(Z)$。

8. 设随机变量 X_1 服从参数为 $\lambda=\frac{1}{2}$ 的指数分布,随机变量 X_2 的概率密度函数为

$$f(x)=\begin{cases} cxe^{-\frac{x}{2}}, & x\geqslant 0 \\ 0, & x>0 \end{cases}$$

试求:(1)$E(X_1)$和 $D(X_1)$;

(2)通过 $E(X_1)$ 和 $D(X_1)$ 计算 c 和 $E(X_2)$ 的值。

9. 设 $X\sim N(1, 2)$，Y 服从参数为 3 的 Poisson 分布，且 X 与 Y 独立，求 $D(XY)$。

10. 掷一颗骰子 1620 次，则“六点”出现的次数 X 的期望和方差为多少？

11. 已知 $X\sim B(n, p)$，且 $E(X)=3$，$D(X)=2$，试求 X 的全部可能取值，并计算 $P(X\leqslant 8)$。

12. 对球的直径作近似测量，其值均匀分布在区间 $[a, b]$ 上，试求球的体积的数学期望。

13. 设连续型随机变量 X 的概率密度为

$$f(x)=\begin{cases} ax^2+bx+c, & 0<x<1 \\ 0, & \text{其它} \end{cases}$$

且 $E(X)=0.5$，$D(X)=0.15$，求系数 a, b, c。

14. 设 X 与 Y 相互独立，其概率密度分别为

$$f_X(x)=\begin{cases} 2x, & 0\leqslant x\leqslant 1 \\ 0, & \text{其它} \end{cases} \qquad f_Y(y)=\begin{cases} e^{-(y-5)}, & y>5 \\ 0, & \text{其它} \end{cases}$$

求 $E(XY)$。

15. 设连续型随机变量 X 在区间 $[1, 2]$ 上服从均匀分布，随机变量

$$Y=\begin{cases} 1, & X>0 \\ 0, & X=0 \\ -1, & X<0 \end{cases}$$

求方差 $D(Y)$。

16. 设某产品每周需求量为 Q，Q 的可能取值为 1,2,3,4,5(等可能取各值)，生产每件产品成本是 $c_1=3$ 元，每件产品售价 $c_2=9$ 元，没有售出的产品以每件 $c_3=1$ 元的费用存入仓库。问生产者每周生产多少件产品可使所期望的利润最大？

17. 在每次实验中事件 A 发生的概率为 0.5，利用切比雪夫不等式估计，在 1000 次独立重复实验中，事件 A 发生的次数在 400～600 之间的概率。

18. 电视台举办智力竞猜，有两种类型的题目：A 类为历史题，B 类为地理题。竞猜者可以自己选择顺序，只有猜对了第一题后猜才有权猜第二题。猜对 A 类题得 a 分，猜对 B 类题得 b 分。现假定某人猜对 A 类题和 B 类题的概率分别为 p 和 q，且此事件是独立的。试问他应当先猜哪类题，可使他的期望得分最高？

19. 已知 $X_i\sim N(0, 1)$，$i=1, 2, 3$，且 X_i 相互独立，令 $\overline{X}=\frac{1}{3}\sum_{i=1}^{3}X_i$，$Y=\sum_{i=1}^{3}(X_i-\overline{X})^2$，求 $E(Y)$。

20. 一辆飞机场的交通车,送 25 名乘客到 9 个站,假设每个乘客都等可能地在任一车站下车,并且他们下车与否相互独立。又知交通车只在有人下车时才停车。求该交通车停车次数的数学期望。

4.2 随机向量的数字特征

1. 随机向量的数学期望

定义 4.2.1 设$(X_1, X_2, \cdots, X_n)$是 n 维随机向量,且每个随机变量 X_i 的期望$E(X_i)$和方差 $D(X_i)$ $(i=1, 2, \cdots, n)$都存在,称$(E(X_1), E(X_2), \cdots, E(X_n))$为$(X_1, X_2, \cdots, X_n)$的**期望向量(expected value vector)**,简称**期望(expectancy)**。称$(D(X_1), D(X_2), \cdots, D(X_n))$为$(X_1, X_2, \cdots, X_n)$的**方差(variance)**。

这一定义不过是一维情况的简单推广,它只是反映了随机向量的各个分量作为一维随机变量的取值情况。但对于各个随机变量之间的联系没有反映出来,而这一点在实际中又是十分重要的。下面我们就二维的情况重点讨论这个问题。

2. 协方差

客观事物是错综复杂的,但这种错综复杂并不是杂乱无章,而是表现为广泛的相互联系上。因此,如何有效地刻画随机变量之间的联系,对于我们研究和控制随机现象是个重要的问题。例如,某种商品的销售量与价格、居民收入、促销手段、市场的饱和度、居民的爱好等许多因素有关,了解销售量与各因素相关程度的高低,对销售商制定销售策略是至关重要的。一个儿童的智商与营养、教育、环境、遗传等许多因素有关,到底哪一个因素起的作用大呢?掌握这些情况,对培养儿童是很重要的。协方差正是刻画随机变量之间联系是否紧密的一个重要的数字特征。

定义 4.2.2 对随机向量(X, Y),若 $E[(X-E(X))(Y-E(Y))]$存在,则称它为 X 和 Y 的**协方差(covariance)**,记为 $\mathrm{cov}(X, Y)$,即

$$\mathrm{cov}(X, Y)=E[(X-E(X))(Y-E(Y))]$$

从上述定义中可见,当 $X=Y$,即它们是同一个随机变量时,有

$$\mathrm{cov}(X, Y)=E[X-E(X)]^2=D(X)$$

此时,协方差就成为该随机变量的方差。因此,可以说方差是协方差的一个特例,而协方差是方差的推广。既然方差反映了随机变量本身的离散程度,那么用协方差反映两个随机变量之间的"离散"程度也就很自然了。

例 4.2.1 设(X, Y)服从二维正态分布,其联合密度为

$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{2\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{2\sigma_2^2}\right)\right\}$$

其中$-\infty<x,\ y<+\infty$，求 $\mathrm{cov}(X,\ Y)$。

解　由题知 $E(X)=\mu_1$，$E(Y)=\mu_2$，$D(X)=\sigma_1^2$，$D(Y)=\sigma_2^2$，于是

$$\begin{aligned}
&\mathrm{cov}(X,\ Y)\\
&=E[(X-E(X))(Y-E(Y))]\\
&=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-\mu_1)(y-\mu_2)f(x,\ y)\mathrm{d}x\mathrm{d}y\\
&=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-\mu_1)(y-\mu_2)\\
&\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{2\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\right)\right\}\mathrm{d}x\mathrm{d}y
\end{aligned}$$

令 $t=\frac{x-\mu_1}{\sigma_1}$，$s=\frac{y-\mu_2}{\sigma_2}$，则

$$\begin{aligned}
&\mathrm{cov}(X,\ Y)\\
&=\frac{\sigma_1\sigma_2}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}ts\cdot\exp\left\{-\frac{1}{2(1-\rho^2)}(t^2-2\rho ts+s^2)\right\}\mathrm{d}t\mathrm{d}s\\
&=\frac{\sigma_1\sigma_2}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty}ts\cdot\exp\left\{-\frac{1}{2(1-\rho^2)}((t-\rho s)^2+(1-\rho^2)s^2)\right\}\mathrm{d}t\right]\mathrm{d}s\\
&=\frac{\sigma_1\sigma_2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}s\cdot\mathrm{e}^{-\frac{s^2}{2}}\left[\frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\int_{-\infty}^{+\infty}t\cdot\mathrm{e}^{-\frac{(t-\rho s)^2}{2(1-\rho^2)}}\mathrm{d}t\right]\mathrm{d}s
\end{aligned}$$

上式中方括号内的积分恰好是正态分布 $N(\rho s,\ 1-\rho^2)$ 的数学期望，因此其积分值为 ρs，于是

$$\begin{aligned}
\mathrm{cov}(X,\ Y)&=\frac{\sigma_1\sigma_2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\rho\cdot s^2\cdot\mathrm{e}^{-\frac{s^2}{2}}\mathrm{d}s\\
&=\rho\sigma_1\sigma_2\int_{-\infty}^{+\infty}s^2\cdot\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{s^2}{2}}\mathrm{d}s\\
&=\rho\sigma_1\sigma_2
\end{aligned}$$

这说明二维正态分布的协方差不仅与其标准差 σ_1，σ_2 有关，而且还与其第五个 ρ 参数有关。

由协方差的定义可以得到一些有关协方差的性质。

性质 1　$\mathrm{cov}(X,\ Y)=E(XY)-E(X)E(Y)$。

性质 2　$\mathrm{cov}(X,\ Y)=\mathrm{cov}(Y,\ X)$。

性质 3　$\mathrm{cov}(aX,bY)=ab\mathrm{cov}(X,\ Y)$，$a,b$ 为常数。

性质 4　$\mathrm{cov}(X_1+X_2,\ Y)=\mathrm{cov}(X_1,\ Y)+\mathrm{cov}(X_2,\ Y)$。

证 这里只证性质 1 和性质 4

（性质 1）

$$\begin{aligned}\operatorname{cov}(X, Y) &= E[(X-E(X))(Y-E(Y))] \\ &= E[XY - XE(Y) - YE(X) + E(X)E(Y)] \\ &= E(XY) - E(X)E(Y)\end{aligned}$$

有了这个性质，随机变量 $X+Y$ 的方差可以写成

$$D(X \pm Y) = D(X) + D(Y) \pm 2\operatorname{cov}(X, Y)$$

由这个性质也可以证明性质 4。

（性质 4）

$$\begin{aligned}\operatorname{cov}(X_1 + X_2, Y) &= E[(X_1 + X_2) \cdot Y] - E(X_1 + X_2)E(Y) \\ &= E[X_1 Y + X_2 Y] - [E(X_1) + E(X_2)]E(Y) \\ &= E(X_1 Y) + E(X_2 Y) - E(X_1)E(Y) - E(X_2)E(Y) \\ &= \operatorname{cov}(X_1, Y) + \operatorname{cov}(X_2, Y)\end{aligned}$$

3. 相关系数

由于协方差的值的大小与随机变量所取的单位有关，这对于评价两个随机变量之间的相依关系的程度大小有时候是不利的。为此，我们在协方差的基础上进行修改，就得到一个与单位无关的更常用的另一指标——相关系数。

定义 4.2.3 对随机向量 (X, Y)，如果 $D(X)$ 和 $D(Y)$ 均存在，并且均不为零，则称

$$\rho_{XY} = \frac{\operatorname{cov}(X, Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

为 X 与 Y 的**相关系数(correlation coefficient)**。

从上述定义中可见 X 与 Y 的相关系数，事实上就是标准化的随机变量 $\frac{X-E(X)}{\sqrt{D(X)}}$ 与 $\frac{Y-E(Y)}{\sqrt{D(Y)}}$ 的协方差。它与随机变量所使用的单位无关，是一个规范化的指标，因此在使用上比协方差更方便。

在例 4.2.1 中，二维正态分布的协方差 $\operatorname{cov}(X, Y) = \rho\sigma_1\sigma_2$，而 $D(X) = \sigma_1^2$，$D(Y) = \sigma_2^2$，所以二维正态分布的相关系数为 $\rho_{XY} = \frac{\operatorname{cov}(X, Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{\rho\sigma_1\sigma_2}{\sigma_1\sigma_2} = \rho$。

例 4.2.2 已知随机变量 X 与 Y 分别服从正态分布 $N(1, 3^2)$ 和 $N(0, 4^2)$，

试讨论①若 $\rho_{XY}=0$,求(X, Y)的联合密度;②若 $\rho_{XY}=-\frac{1}{2}$, $Z=\frac{X}{3}+\frac{Y}{2}$,求$E(Z)$, $D(Z)$和 ρ_{XZ}。

解　①已知 $\rho_{XY}=0$,X 与 Y 相互独立,所以(X, Y)的联合密度为

$$\begin{aligned}f(x, y) &= f_X(x)\cdot f_Y(y)\\&=\frac{1}{3\sqrt{2\pi}}e^{-\frac{(x-1)^2}{2\times 3^2}}\cdot\frac{1}{4\sqrt{2\pi}}e^{-\frac{y^2}{2\times 4^2}}\\&=\frac{1}{24\pi}e^{-\frac{(x-1)^2}{18}-\frac{y^2}{32}}\end{aligned}$$

②因为 $E(X)=1$,$E(Y)=0$,$D(X)=3^2$,$D(Y)=4^2$,于是

$$E(Z)=E\left(\frac{X}{3}+\frac{Y}{2}\right)=\frac{1}{3}E(X)+\frac{1}{2}E(Y)=\frac{1}{3}\times 1+\frac{1}{2}\times 0=\frac{1}{3}$$

$$\begin{aligned}D(Z)&=D\left(\frac{X}{3}+\frac{Y}{2}\right)=\frac{1}{9}D(X)+\frac{1}{4}D(\bar{y})+2\mathrm{cov}\left(\frac{X}{3}, \frac{Y}{2}\right)\\&=\frac{1}{9}D(X)+\frac{1}{4}D(Y)+\frac{1}{3}\rho_{XY}\cdot\sqrt{D(X)}\cdot\sqrt{D(Y)}\\&=\frac{1}{9}\times 3^2+\frac{1}{4}\times 4^2+\frac{1}{3}\times(-\frac{1}{2})\cdot\sqrt{3^2}\cdot\sqrt{4^2}\\&=3\end{aligned}$$

又由于

$$\begin{aligned}\mathrm{cov}(X, Z)&=\mathrm{cov}\left(X, \frac{X}{3}+\frac{Y}{2}\right)=\mathrm{cov}\left(X, \frac{X}{3}\right)+\mathrm{cov}\left(X, \frac{Y}{2}\right)\\&=\frac{1}{3}\mathrm{cov}(X, X)+\frac{1}{2}\mathrm{cov}(X, Y)\\&=\frac{1}{3}D(X)+\frac{1}{2}\rho_{XY}\sqrt{D(X)}\sqrt{D(Y)}\\&=\frac{1}{9}\times 3^2+\frac{1}{2}\times(-\frac{1}{2})\times\sqrt{3^2}\times\sqrt{4^2}\\&=0\end{aligned}$$

所以

$$\rho_{XZ}=\frac{\mathrm{cov}(X, Z)}{\sqrt{D(X)}\sqrt{D(Z)}}=0$$

由相关系数的定义可以得出 ρ_{XY}的有关性质。

性质 1　$|\rho_{XY}\leqslant 1|$

性质 2　如果随机变量 Y 是 X 的线性函数,即 $Y=aX+b$(a, b 为常数, $a\neq 0$),则当 $a>0$ 时,$\rho_{XY}=1$;当 $a<0$ 时,$\rho_{XY}=-1$。

性质 3　$|\rho_{XY}|=1$ 的充要条件是存在常数 a,b,使 $P(Y=aX+b)=1$。

性质 4　若 X 与 Y 相互独立,则 $\rho_{XY}=0$。

在证明性质 1 之前,须先证明一个重要的公式——**柯西(Cauchy)-许瓦兹(Schwarz)不等式**。即若随机向量(X, Y)满足 $E(X^2)$与 $E(Y^2)$都存在,则有

$$[E(XY)]^2\leqslant E(X^2)E(Y^2)$$

证　对任意的实数 t,有

$$f(t)=E[(tX+Y)^2]=t^2E(X^2)-2tE(XY)+E(Y^2)\geqslant 0$$

所以

$$\Delta=4[E(XY)]^2-4E(X^2)E(Y^2)\leqslant 0$$

即

$$[E(XY)]^2\leqslant E(X^2)E(Y^2)$$

下面证明相关系数的性质,只证明性质 1 和性质 2,其它两个性质留给读者证明。

证　(性质 1)由柯西-许瓦兹不等式,有

$$\begin{aligned}[\mathrm{cov}(X, Y)]^2 &= [E(X-E(X))(Y-E(Y))]^2\\ &\leqslant E(X-E(X))^2\cdot E(Y-E(Y))^2\\ &= D(X)D(Y)\end{aligned}$$

所以

$$|\rho_{XY}|=\frac{|\mathrm{cov}(X, Y)|}{\sqrt{D(X)}\sqrt{D(Y)}}\leqslant 1$$

(性质 2)因为

$$\begin{aligned}\mathrm{cov}(X, Y) &= E[(X-E(X))(aX+b-aE(X)-b)]\\ &= aE(X-E(X))^2 = aD(X)\\ D(Y) &= D(aX+b) = a^2D(X)\end{aligned}$$

所以

$$\rho_{XY}=\frac{\mathrm{cov}(X, Y)}{\sqrt{D(X)}\sqrt{D(Y)}}=\frac{aD(X)}{|a|D(X)}=\begin{cases}1, & a>0\\ -1, & a<0\end{cases}$$

由上述性质可以看出,$|\rho_{XY}|$的大小反映了 X 与 Y 线性相关程度的高低。ρ_{XY} 的取值在区间$[-1, 1]$上。根据它取值的情况,我们有如下定义。

定义 4.2.4　设随机变量 X 与 Y 的相关系数为 ρ_{XY}。若 $\rho_{XY}\neq 0$,则称 X 与 Y **相关(correlation)**;当 $\rho_{XY}>0$ 时,称 X 与 Y **正相关(positive correlation)**;当 $\rho_{XY}<0$ 时,称 X 与 Y **负相关(negative correlation)**。特别当 $\rho_{XY}=1$ 时,称 X 与 Y **完全正相关(complete positive correlation)**;当 $\rho_{XY}=-1$ 时,称 X 与 Y **完全负相关**

(**complete negative correlation**)；当$_{XY}=0$ 时，称 X 与 Y **不相关**(**noninteracting**)。

显然，X 与 Y 不相关；$E(XY)=E(X)E(Y)$；$D(X\pm Y)=D(X)+D(Y)$；$\rho_{XY}=0$；$\text{cov}(X, Y)=0$；$D(X+Y)=D(X-Y)$是等价的。

同时，若 X 与 Y 相互独立，那么 X 与 Y 一定不相关。但反之不一定成立(如前面的例 4.1.12 可以说明这一点)。

当(X, Y)为二维正态分布时，X 与 Y 相互独立和 X 与 Y 不相关是等价的。

例 4.2.3　设(X, Y)在矩形区域 $D=\{(x, y)|0\leqslant x\leqslant 1, 0\leqslant y<1\}$上服从均匀分布，试判断 X 与 Y 是否相关？

解　(方法一)

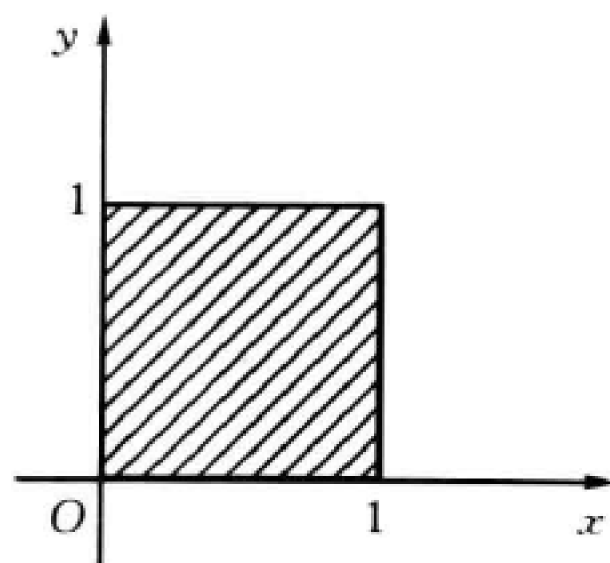


图 4.2.1

因为(X, Y)的联合密度为

$$f(x, y)=\begin{cases}1, & 0\leqslant x\leqslant 1, 0\leqslant y\leqslant 1\\ 0, & \text{其它}\end{cases}$$

所以

$$\begin{aligned}E(XY) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x, y)\mathrm{d}x\mathrm{d}y\\ &= \int_0^1 x\mathrm{d}x \cdot \int_0^1 y\mathrm{d}y\\ &= \frac{1}{4}\end{aligned}$$

$$\begin{aligned}E(X) &= \int_{-\infty}^{+\infty} xf_X(x)\mathrm{d}x = \int_{-\infty}^{+\infty} x\left[\int_{-\infty}^{+\infty} yf(x, y)\mathrm{d}y\right]\mathrm{d}x\\ &= \int_0^1 x\mathrm{d}x = \frac{1}{2}\end{aligned}$$

同理 $E(Y)=\frac{1}{2}$。所以 X 与 Y 的协方差为

$$\text{cov}(X, Y)=E(XY)-E(X)E(Y)=\frac{1}{4}-\frac{1}{2}\times\frac{1}{2}=0$$

即 X 与 Y 不相关。

(方法二)

因为 $f_X(x) = \int_{-\infty}^{+\infty} f(x, y)\mathrm{d}y = \int_0^1 \mathrm{d}y = 1\ (0 \leqslant x \leqslant 1)$，所以

$$f_X(x)=\begin{cases}1, & 0\leqslant x\leqslant 1\\ 0, & \text{其它}\end{cases}$$

同理

$$f_Y(y)=\begin{cases}1, & 0\leqslant y\leqslant 1\\ 0, & \text{其它}\end{cases}$$

于是对任意的 x,y 有

$$f(x,\ y)=f_X(x)f_Y(x)$$

故 X 与 Y 相互独立，X 与 Y 不相关。

例 4.2.4 设$(X,\ Y)$在圆形区域 $D=\{(x,\ y)\mid x^2+y^2\leqslant 1\}$上服从均匀分布，试判断 X 与 Y 是否相关，是否独立？

解 因为$(X,\ Y)$的联合密度为

$$f(x,\ y)=\begin{cases}\dfrac{1}{\pi}, & x^2+y^2\leqslant 1\\ 0, & \text{其它}\end{cases}$$

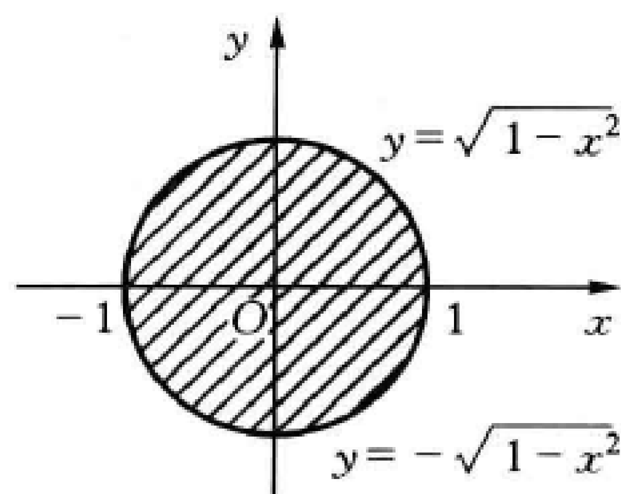


图 4.2.2

所以

$$\begin{aligned}E(XY)&=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xyf(x,\ y)\mathrm{d}x\mathrm{d}y\\&=\int_{-1}^{1}\mathrm{d}x\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}}xy\ \frac{1}{\pi}\mathrm{d}y\\&=0\end{aligned}$$

$$\begin{aligned}E(X)&=\int_{-\infty}^{+\infty}xf_X(x)\mathrm{d}x=\int_{-1}^{1}\mathrm{d}x\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}}x\ \frac{1}{\pi}\mathrm{d}y\\&=\frac{2}{\pi}\int_{-1}^{1}x\ \sqrt{1-x^2}\mathrm{d}x=0\end{aligned}$$

同理 $E(Y)=0$。所以 X 与 Y 的协方差为

$$\mathrm{cov}(X,\ Y)=E(XY)-E(X)E(Y)=0$$

即 X 与 Y 不相关。

因为 $f_X(x)=\int_{-\infty}^{+\infty}f(x,\ y)\mathrm{d}y=\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}}\frac{1}{\pi}\mathrm{d}y=\frac{2}{\pi}\ \sqrt{1-x^2}\ (-1\leqslant x\leqslant 1)$，

所以

$$f_X(x)=\begin{cases}\dfrac{2}{\pi}\sqrt{1-x^2}, & (-1\leqslant x\leqslant 1)\\ 0, & \text{其它}\end{cases}$$

同理

$$f_Y(y)=\begin{cases}\dfrac{2}{\pi}\sqrt{1-y^2}, & (-1\leqslant y\leqslant 1)\\ 0, & \text{其它}\end{cases}$$

由于 $x,\ y$ 对有

$$f(x,\ y)\neq f_X(x)f_Y(x)$$

故 X 与 Y 不独立。

为便于以后应用，我们再给出随机变量的“矩”的概念。对于随机变量 X，若 $E(|X|^k)$（k为正整数）存在，则称 $E(X^k)$ 为 X 的 k 阶原点矩，特别一阶原点矩就是数学期望（均值）。称 $E((X-E(X))^k$ 为 X 的 k 阶中心矩，特别二阶中心矩就是方差。所以，矩的概念是均值和方差概念的推广。

练习 4.2

1. 设(X, Y)的联合概率分布为

X \ Y	-1	0	1
-1	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
0	$\frac{1}{8}$	0	$\frac{1}{8}$
1	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

求：(1) (X, Y)的期望与方差；(2)$\mathrm{cov}(X, Y)$与 ρ_{XY}。

(3)问 X 与 Y 是否相关，是否独立?

2. 设(X, Y)的联合密度为

$$f(x, y)=\begin{cases}Cxy, & 0\leqslant x\leqslant 1,\ 0\leqslant y\leqslant x\\ 0, & \text{其它}\end{cases}$$

求：(1) 常数 C；　　(2) (X, Y)的期望与方差

(3) $\mathrm{cov}(X, Y)$与 ρ_{XY}；　　(4) 判断 X 与 Y 是否相关，是否独立?

3. 设 $D(X)=25$，$D(Y)=36$，$\rho_{XY}=0.4$，求 $D(X+Y)$和 $D(X-Y)$。

4. 设 a, b, c, d 为常数，X 与 Y 的相关系数为 ρ_{XY}，试求 $Z_1=aX+b$，$Z_2=cY+d$的相关系数 ρ。

5. 设 X 服从参数为 2 的 Poisson 分布，$Y=3X-2$，试求 $E(Y)$，$D(Y)$，$\mathrm{cov}(X, Y)$及 ρ_{XY}。

6. 设 X 服从参数为 1 的指数分布，$Y=X+\mathrm{e}^{-2X}$，试求 $E(Y)$，$D(Y)$，$\mathrm{cov}(X, Y)$及 ρ_{XY}。

7. 设随机变量 X 与 Y 相互独立且都服从正态分布 $N(\mu, \sigma^2)$，试求 $Z_1=\alpha X+\beta Y$，$Z_2=\alpha X-\beta Y$ 的相关系数 $\rho_{Z_1Z_2}$，α，β 其中为常数。

8. 设随机变量 X 与 Y 均服从标准正态分布 $N(0, 1)$，它们的相关系数为 $\rho_{XY}=\dfrac{1}{2}$，$Z_1=aX$，$Z_2=bX+cY$，试求 a, b, c 的值，使 $D(Z_1)=D(Z_2)=1$，且 Z_1

与 Z_2 不相关。

9. 设随机变量 X 与 Y 的数学期望分别为-2和 2,方差分别为 1 和 4,而相关系数为-0.5,根据切比雪夫不等式估计 $P\{|X+Y|\geq 6\}$。

4.3 大数定律与中心极限定理

1. 大数定律

在日常生活中,评判一个事物的好坏往往要进行多次的观察和记录,在条件允许的情况下,用多次结果的平均作为衡量的标准,而且随着实验次数的增加这种平均越来越接近于该事物的真实状况。例如,在运动会上体操运动员某动作的成绩,是将各个评委打的分数加以平均作为最终成绩,而且参评的评委越多,这个平均分应越接近于运动员的真实成绩;在评估一个学校的某科成绩时,随机选出 50 名学生参加考试,这 50 名学生的平均成绩就作为评估该校水平的标准,而且参加考试的学生越多,这平均成绩就越接近于该校的真实水平;在实验中要测量某一物体的长度,一般要进行多次测量,然后将多次测量的结果加以平均作为该物体的长度;而且随着测量次数的增加,该平均值越来越接近于该物体长度的真实值。

以上事例说明,一次实验的结果,可能由于种种随机因素而产生一些波动,偏离其本质。但是作为大量实验的平均结果,事件发生的频率和大量观测值的平均值都是有独立于人的主观意志的稳定性,这种稳定性就为大数定律提供了丰富的生活背景。

定义 4.3.1 设随机变量序列$\{X_n\}$,如果存在一个常数 a,使得对任意的 $\varepsilon>0$,有

$$\lim_{n\to\infty}P\{|X_n-a|<\varepsilon\}=1$$

则称$\{X_n\}$**依概率收敛于** a。记作 $X_n\xrightarrow{p}a$。

依概率收敛是概率论中特有的一种思维方式。随机变量序列在无限逼近数 a 时,可能会出现波动,有时波动可能会很大,只是这样的机会概率越来越少。这种复杂的收敛方式在生活中往往比微积分中的收敛方式要多。

定理 4.3.1(切比雪夫大数定律) 设随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$相互独立,且均存在数学期望 $E(X_n)=\mu_n$,方差 $D(X_n)=\sigma_n^2<k$ $(n=1, 2, \cdots)$,其中常数 k 与 n 无关,则对任意的 $\varepsilon>0$,有

$$\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}\mu_i\right|<\varepsilon\right\}=1$$

证 由于 $X_1, X_2, \cdots, X_n, \cdots$相互独立,所以

$$E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\frac{1}{n}\sum_{i=1}^{n}\mu_i$$

$$D\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}D(X_i)=\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2<\frac{k}{n}$$

由切比雪夫不等式，有

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}\mu_i\right|<\varepsilon\right\}\geqslant 1-\frac{1}{\varepsilon^2}\cdot\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2>1-\frac{k}{n\varepsilon^2}$$

而

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}\mu_i\right|<\varepsilon\right\}\leqslant 1,\ \lim_{n\to\infty}\left(1-\frac{k}{n\varepsilon^2}\right)=1$$

所以

$$\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}\mu_i\right|<\varepsilon\right\}=1$$

由切比雪夫大数定律可以看出，无论随机现象的个别结果如何，大量相互独立随机现象的平均结果实际上不受随机现象个别结果的影响，平均后的随机变量 $\frac{1}{n}\sum_{i=1}^{n}X_i$ 将比较紧密地聚集在它的期望 $\frac{1}{n}\sum_{i=1}^{n}\mu_i$ 的附近，它与期望的距离依概率收敛于零（$n\to\infty$）。

特别地，当 $E(X_i)=\mu,\ D(X_i)=\sigma^2\ (i=1,\ 2,\ \cdots)$ 时，将得到下面应用更广泛的结论。

定理 4.3.2　设 $X_1,\ X_2,\ \cdots,\ X_n,\ \cdots$ 为相互独立的随机变量序列，且有相同的期望与方差：$E(X_i)=\mu,\ D(X_i)=\sigma^2\ (i=1,\ 2,\ \cdots)$，则对任意的 $\varepsilon>0$，有

$$\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right|<\varepsilon\right\}=1$$

这一结论说明所有观察结果的算术平均值依概率收敛于期望值——被观察值的真值。这就为我们实际生活中用算术平均值代替真值提供了理论依据。同时，这一推论也是后面统计中用样本矩去估计总体矩的理论依据。

定理 4.3.3（贝努里大数定律）　设每次实验中事件 A 发生的概率为 p，n 次重复独立实验中事件 A 发生的次数为 μ_n，则对任意的 $\varepsilon>0$，事件 A 的频率 $\frac{\mu_n}{n}$ 有

$$\lim_{n\to\infty}P\left\{\left|\frac{\mu_n}{n}-p\right|<\varepsilon\right\}=1$$

证　设随机变量服从 0－1 分布，即

$$X_i=\begin{cases}0, & \text{在第 } i \text{ 次试验中 } A \text{ 不发生}\\ 1, & \text{在第 } i \text{ 次试验中 } A \text{ 发生}\end{cases}$$

$i=1, 2, \cdots, n$。显然有

$$\mu_n = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i$$

又 $E(X_i)=p, D(X_i)=p(1-p), i=1, 2, \cdots n$,且 $X_1, X_2, \cdots, X_n$ 相互独立,由定理 4.3.2 知

$$\lim_{n\to\infty} P\left\{ \left| \frac{\mu_n}{n} - p \right| < \varepsilon \right\} = 1$$

贝努里大数定律告诉我们,当实验在不变的条件下重复进行多次时,事件 A 的频率依概率收敛于随机事件 A 的概率 p。这就为我们推测概率值提供了一种简便易行的方法,即用频率代替概率。同时也使概率这一概念的背景更加清晰可靠。

例 4.3.1　新生儿的体重是个随机变量 X。我们随机地抽取了 2006 年 3 月份新生儿(男)n 名,测其体重得到 n 个数据 $x_1, x_2, \cdots, x_n$。令 $\mu_n(x)$ 代表 $x_1, x_2, \cdots, x_n$ 中小于或等于 x 的个数,那么

$$F_n^*(x) = \frac{1}{n}\mu_n(x)$$

就代表了体重 X 小于 x 的频率,我们称它为 X 的经验分布函数(也是累计频率函数)。它是事件($X \leqslant x$)的频率,从而是概率 $P(X \leqslant x)$ 的近似值。由大数定律可知,当 n 越来越大时,对每一点 x,经验分布函数 $F_n^*(x)$ 都会趋于随机变量 X 的分布函数 $F(x)$。即

$$\lim_{n\to\infty} P\left\{ \left| \frac{\mu_n}{n} - P(X \leqslant x) \right| < \varepsilon \right\} = 1$$

$F_n^*(x)$ 依概率收敛于 $F(x)$,因此大数定律也是实际应用中用经验分布代替理论分布的理论基础。

2. 中心极限定理

在实际生活中我们遇到的随机变量大多数都服从正态分布或近似正态分布。例如,上例中某地区 2006 年 3 月份新生儿的身高和体重服从正态分布;工厂中产品某性能的指标一般也服从正态分布,因此生产过程中的质量控制图就是以正态曲线为基础绘制的;在对经济问题(如商品价格、商品销售量)进行定量分析时,往往在找出主要的因素之外,其它各种因素的综合影响就用一个服从正态分布的随机变量来表示等等。这些随机变量有一个共同的特点:由若干相互独立的随机变量的和构成。事实上,如果一个随机现象由众多的随机因素所引起,每一因素在总的变化里作用不显著。可以推断,不管每一微小因素原来服从什么分布,描述这个随机现象的随机变量都近似地服从正态分布。

定理 4.3.4　独立同分布中心极限定理（**independent identically distribution central limit theorem**）设随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立同分布，而且 $E(X_i)=u,\ D(X_i)=\sigma^2\ (i=1, 2, \cdots),\ \sigma^2>0$，则随机变量

$$Y_n=\frac{\sum_{i=1}^{n}X_i-n\mu}{\sqrt{n}\sigma}$$

的分布函数 $F_{Y_n}(x)$ 对任意 x 有

$$\lim_{n\to\infty}F_{Y_n}=\lim_{n\to\infty}P\left(\frac{\sum_{i=1}^{n}X_i-n\mu}{\sqrt{n}\sigma}\leqslant x\right)=\Phi(x)$$

这个定理表明只要 n 比较大，随机变量 Y_n 就近似地服从标准正态分布。因而 $\sum_{i=1}^{n}X_i$ 近似地服从正态分布 $N(n\mu,\ n\sigma^2)$。中心极限定理确立了正态分布在各种分布中的首要地位，它也是数理统计中大样本处理方法必不可少的理论基础。

若记 $\beta=\Phi(x)$，则由中心极限定理给出的近似公式 $P(Y_n\leqslant x)\approx\Phi(x)=\beta$ 可用来解决三类计算问题：(1)已知 n, x 求 β；(2)已知 n, β 求 x；(3)已知 x, β 求 n。

例 4.3.2　一生产线生产的产品成箱包装，每箱的重量是随机的。假设每箱平均重 50 千克，标准差为 5 千克。若用最大载重量为 5 吨的汽车承运，试用中心极限定理说明每车最多可以装多少箱，才能保障不超载的概率大于 0.977。

解　设 $X_i(i=1, 2, \cdots, n)$ 为装运的第 i 箱的重量（单位：千克），n 是所求的箱数。由题意可把 $X_1, X_2, \cdots, X_n$ 看作独立同分布的随机变量，令

$$Y_n=X_1+X_2+\cdots+X_n=\sum_{i=1}^{n}X_i$$

那么 Y_n 就是这 n 箱货物的总重量。又知 $E(X_i)=50, D(X_i)=5^2, E(Y_n)=50n, D(Y_n)=25\sqrt{n}$。由中心极限定理有

$$P(Y_n\leqslant 5000)\approx\Phi\left(\frac{5000-50n}{5\sqrt{n}}\right)>0.977=\Phi(2)$$

所以

$$\frac{1000-10n}{\sqrt{n}}>2,\quad n<98.0199$$

最多可以装 98 箱。

例 4.3.3　设 $X_1^2, X_2^2, \cdots, X_n^2, \cdots$ 是独立同分布的随机变量序列，其中 $X_i\sim N(0, 1)\ (i=1, 2, \cdots)$，那么令

$$Y_n=\frac{\sum_{i=1}^{n}X_i^2-n}{\sqrt{2n}}$$

则 Y_n 渐近服从标准正态分布。

证 显然依题知

$$E(X_i^2)=D(X_i)-E^2(X_i)=1$$

$$D(X_i^2)=E(X_i^4)-E^2(X_i^2)$$

$$=\int_{-\infty}^{+\infty}x^4\ \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}\mathrm{d}x-1=3-1=2\ (i=1,2,\cdots),$$

所以

$$E\left(\sum_{i=1}^{n}X_i^2\right)=n,\quad D\left(\sum_{i=1}^{n}X_i^2\right)=2n$$

由中心极限定理有

$$\lim_{n\to\infty}P\left(\frac{\sum_{i=1}^{n}X_i^2-n\mu}{\sqrt{2n}}\leqslant x\right)=\Phi(x)$$

即 Y_n 渐近服从标准正态分布。

在数理统计中，随机变量 $\sum_{i=1}^{n}X_i^2$ 的分布是一个重要的分布（χ^2 -分布）。上例告诉我们，当 $n\to\infty$ 时，$\sum_{i=1}^{n}X_i^2$ 渐近服从正态分布 $N(n,2n)$。

下面再介绍一个定理，它实际上也是定理 4.3.4 的一个自然推论。

定理 4.3.5 设随机变量 $\mu_n(n=1,2,\cdots)$ 服从 $B(n,p)$，则对任意 x，有

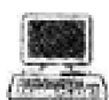
$$\lim_{n\to\infty}P\left(\frac{\mu_n-np}{\sqrt{np(1-p)}}\leqslant x\right)=\Phi(x)$$

因为二项分布为离散型分布，而正态分布为连续型分布，所以用正态分布作为二项分布的近似计算中，作些修正可以提高精度，若 $a<b$，均为整数，一般先作如下修正后再用正态近似

$$P(a\leqslant\mu_n\leqslant b)=P(a-0.5<\mu_n<b+0.5)\approx\Phi\left(\frac{b+0.5-np}{\sqrt{np(1-p)}}\right)-\Phi\left(\frac{a-0.5-np}{\sqrt{np(1-p)}}\right)$$

证明从略。

这个定理说明，当 $n\to\infty$时，二项分布的极限分布为正态分布。它给出了二项分布的一种近似计算方法。我们知道，在 n 次重复独立实验中，事件 A 出现的次数 μ_n 是服从二项分布的，因此当 n 较大时，关于 μ_n 的计算均可用正态分布去近似。

 **中心极限定理演示实验**

在这里我们也可以看出，大数定律定性地指出了许多独立的随机因素叠加的平均结果收敛于常数而趋于稳定。而中心极限定理则定量地给出了这个叠加结果的极限分布。从而使对随机现象的研究更加具体、全面，更有操作性。

例 4.3.4　设电路供电网中有 10000 盏灯，夜晚每一盏灯开着的概率都是 0.7，假定各灯开、关时间彼此无关，计算同时开着的灯数大于 6800 且小于 7200 的概率。

解　设 X 表示同时开着的灯数，则依题知 $X \sim B(10000, 0.7)$。显然 $E(X)=np=7000$，$D(X)=\sqrt{np(1-p)}=45.83$。因此由定理 4.3.5 有

$$\begin{aligned} P(6800 < X < 7200) &\approx \Phi\left(\frac{7200-7000}{45.83}\right)-\Phi\left(\frac{6800-7000}{45.83}\right) \\ &= \Phi(4.36)-\Phi(-4.36) \\ &= 2\Phi(4.36)-1 \\ &= 0.99999 \end{aligned}$$

例 4.3.5　设有 40 个电子元件 $D_1, D_2, \cdots, D_{40}$，它们的使用情况如下：D_1 损坏，D_2 立即使用；D_2 损坏，D_3 立即使用；……。设每个电子元件 D_i 的寿命 X_i 是相互独立同分布的随机变量，$E(X_i)=10$，$D(X_i)=10$，令 X 表示 40 个电子元件使用的总寿命，求 $P(420<X<460)$。

解　显然 $X=\sum\limits_{i=1}^{40}X_i$，由于 $X_1, X_2, \cdots, X_{40}$ 独立同分布，所以

$$E(X)=\sum_{i=1}^{40}E(X_i)=400,\quad D(X)=\sum_{i=1}^{40}D(X_i)=400$$

由定理 4.3.4 知

$$\begin{aligned} P(420 < X < 460) &\approx \Phi\left(\frac{460-400}{20}\right)-\Phi\left(\frac{420-400}{20}\right) \\ &= \Phi(3)-\Phi(1)=0.9986-0.8413 \\ &= 0.1573 \end{aligned}$$

例 4.3.6　分别利用切比雪夫不等式及中心极限定理估计概率 $P\left(\left|\frac{\mu_n}{n}-p\right|\geqslant\varepsilon\right)$，其中 μ_n 是 n 次贝努里试验中事件 A 发生的次数，p 为事件 A 在每次试验中发生的概率，并就 $n=600$，$p=\frac{1}{6}$，$\varepsilon=0.02$ 时进行比较。

解　由于 $\mu_n \sim B(n, p)$，故 $E\left(\frac{\mu_n}{n}\right)=p$，$D\left(\frac{\mu_n}{n}\right)=\frac{p(1-p)}{n}$，由切比雪夫不等式有

$$P\left(\left|\frac{\mu_n}{n}-p\right|\geqslant\varepsilon\right)\leqslant\frac{1}{\varepsilon^2}\cdot\frac{p(1-p)}{n}=\frac{p(1-p)}{n\varepsilon^2}$$

由中心极限定理有

$$\begin{aligned}P\left(\left|\frac{\mu_n}{n}-p\right|\geqslant\varepsilon\right)&=1-P\left(\left|\frac{\mu_n}{n}-p\right|<\varepsilon\right)\\&=1-P\{n(p-\varepsilon)<\mu_n<n(p+\varepsilon)\}\\&\approx1-\left\{\Phi\left(\frac{n(p+\varepsilon)-np}{\sqrt{np(1-p)}}\right)-\Phi\left(\frac{n(p-\varepsilon)-np}{\sqrt{np(1-p)}}\right)\right\}\\&=1-\Phi\left(\frac{n\varepsilon}{\sqrt{np(1-p)}}\right)+\Phi\left(\frac{-n\varepsilon}{\sqrt{np(1-p)}}\right)\\&=2\left[1-\Phi\left(\frac{n\varepsilon}{\sqrt{np(1-p)}}\right)\right]\end{aligned}$$

当 $n=600$，$p=\frac{1}{6}$，$\varepsilon=0.02$ 时，由切比雪夫不等式有

$$P\left(\left|\frac{\mu_n}{n}-p\right|\geqslant\varepsilon\right)\leqslant\frac{p(1-p)}{n\varepsilon^2}=0.5787$$

由中心极限定理得

$$P\left(\left|\frac{\mu_n}{n}-p\right|\geqslant\varepsilon\right)\approx2\left[1-\Phi\left(\frac{n\varepsilon}{\sqrt{np(1-p)}}\right)\right]=0.1886$$

可见，由中心极限定理计算的结果要比用切比雪夫不等式精确得多。

练习 4.3

1. 一个螺丝钉重量是一个随机变量，期望值是 1 两，标准差是 0.1 两，求一盒(100 个)同型号螺丝钉的重量超过 10.2 斤的概率。

2. 对敌人的防御地段进行 100 次轰炸，每次轰炸命中目标的炸弹数目是一个随机变量，其期望值为 2，方差为 1.69，求 100 次轰炸中有 180 颗到 220 颗炸弹命中目标的概率。

3. 在进行加法运算时，为简便起见，每个加法取整数(按四舍五入取最为接近的整数)。可以认为各个加数的取整误差是相互独立的，并且它们都服从[−0.5,0.5]上的均匀分布。求

(1)将 300 个数相加，误差总和的绝对值超过 15 的概率；

(2)最多几个数加在一起，其误差总和的绝对值小于 10 的概率不小于 90%；

(3)如果有 300 个数相加，以 99.7% 的概率断定其误差总和所在的范围。

4. 设有 30 个电子器件，它们的使用寿命 T_1，T_2，…，T_{30} 都服从 $\lambda=0.1$ (单

位:(小时)$^{-1}$)的指数分布。其使用情况是第一个损坏,第二个立即使用;第二个损坏,第三个立即使用;……,令 T 为 30 个器件使用的总计时间,计算 T 超过 360 小时的概率。

5. 某微机系统有 120 个终端,每个终端有 5%的时间在使用。若各终端使用与否是相互独立的,试求有不少于 10 个终端在使用的概率。

6. 设在 n 次贝努里试验中,每次实验事件 A 出现的概率均为 0.7,要使事件 A 出现的频率在 0.68 到 0.72 之间的概率不少于 0.90,问至少要进行多少次试验?

(1) 用切比雪夫不等式估计;(2)用中心极限定理计算。

7. 一保险公司有 10000 人投保,每人每年付 12 元保险费。已知一年内投保人死亡率为 0.006,如死亡,公司付给死者家属 1000 元。求

(1)保险公司年利润为 0 的概率;

(2)保险公司年利润不少于 60000 元的概率。

8. 某车间同型号机床有 200 部,每部机床开动的概率为 0.7。假定各机床开动与否互不影响,开动时每部机床需耗电能 15 个单位。问至少供应多少单位电能才能以 95%的概率保证不致因供电不足而影响生产。

9. 抽样检查产品质量时,如果发现有多于 10 个的次品,则拒绝接受这批产品。设某批产品的次品率为 10%,问至少应抽取多少个产品检查,才能保证拒绝接受该产品的概率达到 0.9?

第 5 章　统计估值

前四章内容已经告诉我们，随机现象可以用随机变量描述，而对随机变量的刻画最好是知道它的分布函数，或至少知道它的某些数字特征. 问题在于，当我们刚接触某一具体的随机现象时，描述这种随机现象的随机变量的分布函数及其数字特征一般来说都是未知的，因此，研究这一随机现象，首先必须解决的问题是如何确定相应的随机变量分布函数或它的某些数字特征？比如，如何确定某地区人均年消费额的分布？如何确定某企业生产的一批电子组件平均使用寿命？普查或逐一地试验固然是获得此类问题最准确结果的方法，但实施这种方法，必然耗费大量的人力、财力，更何况有些试验或费时（如对城市居民市场购买力的调查）或难以办到（如电子组件的使用寿命、轮胎的行使里程等都是破坏性的试验以及因居住、危险而难以接触人群的调查），因此，这种方法在实际中不可行。在长期的实践研究中，人们总结出一种实用而合理的方法 —— 随机抽样法，即从所研究的对象中任意抽取一小部分进行试验或观察，并对所得资料（带有随机误差）加以整理和分析，依据这些资料显示的统计规律性，应用概率论原理，对所研究对象整体的统计规律性作出推断。简言之，由部分推断整体，这就是数理统计的研究方法，其方法的科学体系构成是以概率论为基础的重要应用性学科 —— 数理统计学。随着计算机科学技术的迅速普及，数理统计学在经济、管理、教育、医学、国防、体育、工业、农业、社会等众多学科研究领域中已得到了广泛应用。

5.1　数理统计学中的基本概念

1. 基本概念

为研究方便，先作一些规定。

(1) 总体和个体

我们把研究对象的全体称为**总体(population)**，把构成总体的每一个对象称为**个体(individual)**。比如，要研究某一批电子组件的使用寿命，记寿命为 X，它是一个随机变量，X 的全部取值的集合即为总体，而 X 的每一个可能取值即为个体。

(2)样本和简单随机抽样

我们从总体 X 中随机地抽取 n 个个体,并用 $X_1, X_2, \cdots, X_n$ 表示,这 n 个个体称为取自总体 X 的**样本(sample)**,样本中个体的数目 n 称为**样本容量(size of a sample)**。从总体中随机抽取样本的过程称为**抽样(sampling)**,满足下面两个要求的抽样称为**简单随机抽样(simple random sampling)**。一是样本具有代表性:总体中每个个体被抽到的机会均等,个体的分布能代表总体的分布,即 X_i 与 X 同分布;二是样本具有独立性:总体中每抽出一个个体后,总体的元素个数"几乎"没变,这只要要求样本仅占总体的很小部分,从而保证了 $X_1, X_2, \cdots, X_n$ 的相互独立性. 由简单随机抽样得到的样本称为**简单随机样本(simple random sampling)**,以后如无特别说明,所谈样本均指简单随机样本。样本在抽样前(做一般性研究)为 n 个随机变量,在抽样后(做具体性研究)为 n 个样本值 $x_1, x_2, \cdots, x_n$。

(3)统计量和抽样分布

为了集中样本所带来的总体的信息,我们经常会根据问题的需要,构造有关样本的函数,以便利用这些函数对总体的分布或数字特征作出推断。我们把不含任何未知参数的样本的函数称为**统计量(statistic)**,统计量本身是一个随机变量,且可由样本值计算出来. 统计量的分布称为**抽样分布(sample distribution)**。常用的统计量如下。

样本 k 阶原点矩:$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k (i = 1, 2, \cdots)$,特别样本一阶原点矩 A_1 称为**样本均值(sample mean)**,记为 $\overline{X}$。

样本 k 阶中心矩:$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k (i = 1, 2, \cdots)$,特别"修正样本二阶中心矩" $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ **称为样本方差(sample variance)**,记为 S^2,其算术平方根 S 称为样本标准差。

另外,注意到 $\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$ 是常用的简算公式。

例 5.1.1 设 $X_1, X_2, \cdots, X_n$ 为取自某总体 X 的样本,则 $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ 是(　　)。

(A)样本二阶原点矩　　(B)样本二阶中心矩

(C)统计量　　(D)样本标准差

解　按照定义,应选 C。

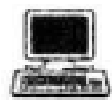
2. 正态总体下的常用统计量及其分布

设总体 $X \sim N(\mu, \sigma^2)$，X_1，X_2，…，X_n 为取自该总体 X 的样本。

(1)四大分布及其分位数

① 标准正态分布 $N(0, 1)$及其上侧分位数。当 μ，σ^2 均已知时，统计量 $Z=\dfrac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$，又设 $0<\alpha<1$，若 $P(Z>Z_\alpha)=\alpha$，则称 Z_α 为 $N(0, 1)$的上侧 α 分位数。易见，$\Phi(Z_\alpha)=1-\alpha$。

② χ^2分布及其上侧分位数。当 $\mu=0$，$\sigma^2=1$ 时，统计量 $\chi^2=\sum\limits_{i=1}^{n} X_i^2$ 称为自由度是 n 的 χ^2 统计量，它所服从的分布称为自由度是 n 的 χ^2 分布，记作 $\chi^2(n)$。易证，χ^2 分布具有可加性，即若 $X \sim \chi^2(m)$，$Y \sim \chi^2(n)$，X，Y 独立，则 $X+Y \sim \chi^2(m+n)$。对于 $0<\alpha<1$，若 $P(\chi^2 > \chi_\alpha^2(n))=\alpha$，则称 $\chi_\alpha^2(n)$ 为自由度是 n 的 χ^2 分布的上侧 α 分位数。

 **χ^2 分布演示实验**

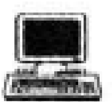
例 5.1.2　设 X_1，X_2，X_3，X_4 是取自总体 $X \sim N(0, 4)$的样本，若 $a(X_1-2X_2)^2+b(3X_3-4X_4)^2$ 服从 $\chi^2(2)$，则常数 $a=$________，$b=$________。

解　因为 $X_1-2X_2 \sim N(0, 20)$，$3X_3-4X_4 \sim N(0, 100)$，所以 $\frac{1}{20}(X_1-2X_2)^2 \sim \chi^2(1)$，$\frac{1}{100}(3X_3-4X_4)^2 \sim \chi^2(1)$，于是按照规定，$\frac{1}{20}(X_1-2X_2)^2+\frac{1}{100}(3X_3-4X_4)^2 \sim \chi^2(2)$，通过比较，易得 $a=\frac{1}{20}$，$b=\frac{1}{100}$。

③ t 分布及其上侧分位数。设 $X \sim N(0, 1)$，$Y \sim \chi^2(n)$，X，Y 相互独立，则称 $T=\dfrac{X}{\sqrt{\frac{Y}{n}}}$ 所服从的分布为自由度是 n 的 t 分布，记作 $t(n)$。

不同于 $\chi^2(n)$的密度曲线的非对称性，$t(n)$的密度曲线类似于 $N(0, 1)$的密度曲线，呈对称性，在实际应用中，当 n 较大时，$t(n)$可用 $N(0, 1)$替代。

对于 $0<\alpha<1$，若 $P(T>t_\alpha(n))=\alpha$，则称 $t_\alpha(n)$为 $t(n)$的上侧 α 分位数。

 **t 分布演示实验**

例 5.1.3　设 $t_\alpha(n)$ 为 $t(n)$ 的上侧 α 分位数，则

$P(T<t_\alpha(n))=$________，$P(T<-t_\alpha(n))=$________，$P(|T|>t_\alpha(n))=$________。

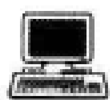
解　依次应填 $1-\alpha$，α，2α。

④ F 分布及其上侧分位数。设 $X\sim\chi^2(n_1)$，$Y\sim\chi^2(n_2)$，X，Y 相互独立，则称 $F=\dfrac{\frac{X}{n_1}}{\frac{Y}{n_2}}$ 所服从的分布为自由度是 (n_1, n_2) 的 F 分布，记作 $F(n_1, n_2)$。易见 $\dfrac{1}{F}\sim F(n_2, n_1)$。对于 $0<\alpha<1$，若 $P(F>F_\alpha(n_1, n_2))=\alpha$，则称 $F_\alpha(n_1, n_2)$ 为 $F(n_1, n_2)$ 的上侧 α 分位数。由 $1-\alpha=P(F>F_{1-\alpha}(n_1, n_2))=P\left(\dfrac{1}{F}<\dfrac{1}{F_{1-\alpha}(n_1, n_2)}\right)=1-P\left(\dfrac{1}{F}>\dfrac{1}{F_{1-\alpha}(n_1, n_2)}\right)$ 知 $P\left(\dfrac{1}{F}>\dfrac{1}{F_{1-\alpha}(n_1, n_2)}\right)=\alpha$，又 $\dfrac{1}{F}\sim F(n_2, n_1)$，所以 $\dfrac{1}{F_{1-\alpha}(n_1, n_2)}=F_\alpha(n_2, n_1)$，即有 $F_{1-\alpha}(n_1, n_2)=\dfrac{1}{F_\alpha(n_2, n_1)}$。

 **F 分布演示实验**

(2) 抽样分布基本定理

设总体 $X\sim N(\mu, \sigma^2)$，X_1，X_2，…，X_n 为取自该总体的样本，则有以下结论成立：① $\overline{X}\sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$；② $\dfrac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$；③ $\overline{X}$ 与 S^2 相互独立。

 **样本均值分布演示实验和样本均值与样本方差独立性演示实验**

由此得下面重要结论：

$$\sqrt{n}\frac{\overline{X}-\mu}{S}\sim t(n-1)$$

又设 X_1，X_2，…，X_{n_1} 为取自总体 $X\sim N(\mu_1, \sigma_1^2)$ 的样本，Y_1，Y_2，…，Y_{n_2} 是取自总体 $Y\sim N(\mu_2, \sigma_2^2)$ 的样本，且两组样本相互独立，则可证明：

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}\sim N(0, 1)$$

$$\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \sim F(n_1-1,\ n_2-1)$$

当 $\sigma_1^2=\sigma_2^2$ 时，记 $S_p^2=\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$，则可证明：

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$$

例 5.1.4　设总体 $X \sim N(0,\ 1)$，X_1，X_2，…，X_n 是取自该总体的样本，则下列各式正确的是(　　)

(A) $\overline{X} \sim N(0,1)$　　　　(B) $n\overline{X} - N(0,\ 1)$

(C) $\dfrac{\overline{X}}{S} \sim t(n-1)$　　　　(D) $\sum\limits_{i=1}^{n} X_i^2 \sim \chi^2(n)$

解　应选 D。

3. SPSS 实现

(1) 利用 SPSS 计算样本均值、样本方差、样本标准差

例 5.1.5　随机抽取 13 位同学的微积分、英语、计算机课程成绩(见表 5.1.1 中 SPSS 数据文件例 5.1.5)，试计算其每门课的样本均值、样本方差、样本标准差。

解　首先在 SPSS 变量编辑窗口定义变量：x 为微积分成绩，y 为英语成绩，z 为计算机成绩，在数据编辑窗口输入数据，得到如表 5.1.1 所示数据文件。

表 5.1.1　随机抽取的 13 位同学的三门课程成绩

Untitled - SPSS Data Editor

File　Edit　View　Data　Transform　Analyze　SPSSAddins

12 :

	x	y	z	var
1	80	66	77	
2	90	78	78	
3	98	88	86	
4	78	78	68	
5	78	78	69	
6	67	79	96	
7	78	98	93	
8	86	87	69	
9	76	87	87	
10	66	86	77	
11	89	85	68	
12	88	67	87	
13	78	76	89	

然后依次选项：Analyze→Descriptive Statistics→Descriptives，即如图 5.1.1 显示。

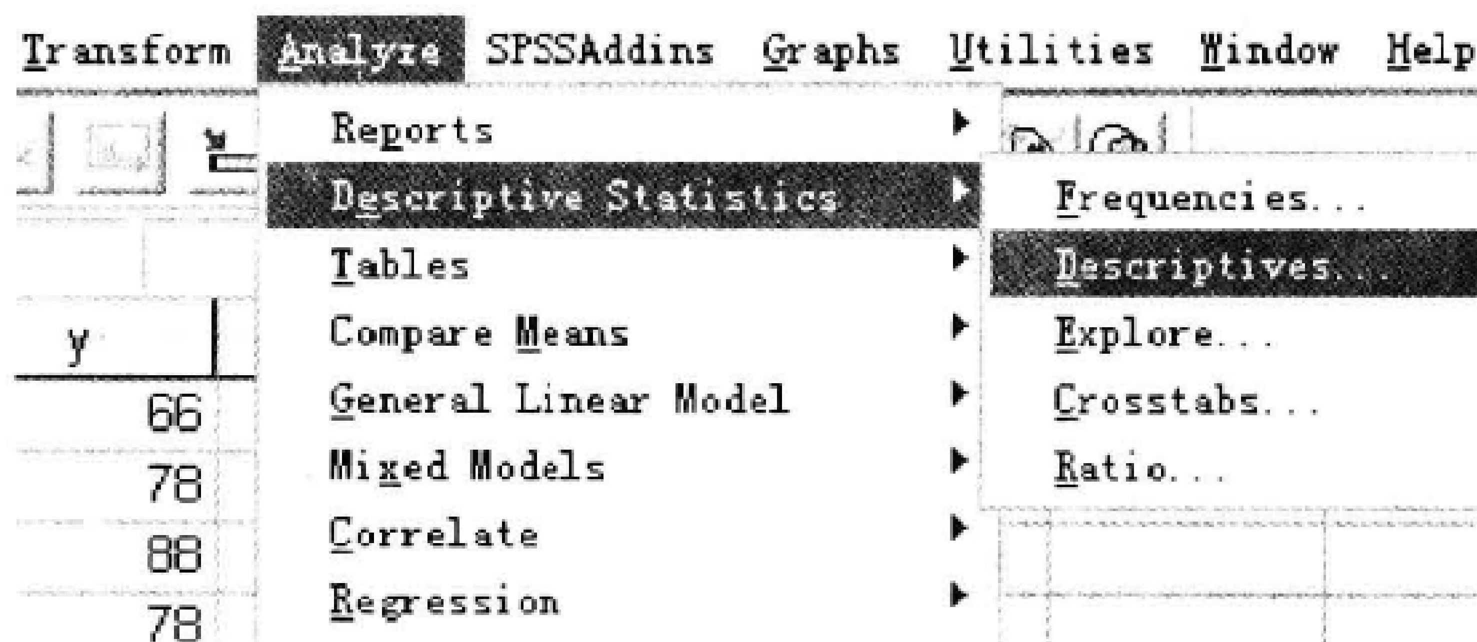


图 5.1.1

在打开的对话框中将源变量栏（左侧）的 x，y，z 选入分析变量栏，点击 Options...按钮，在弹出的子对话框中选择相应的选项（如图 5.1.2）。

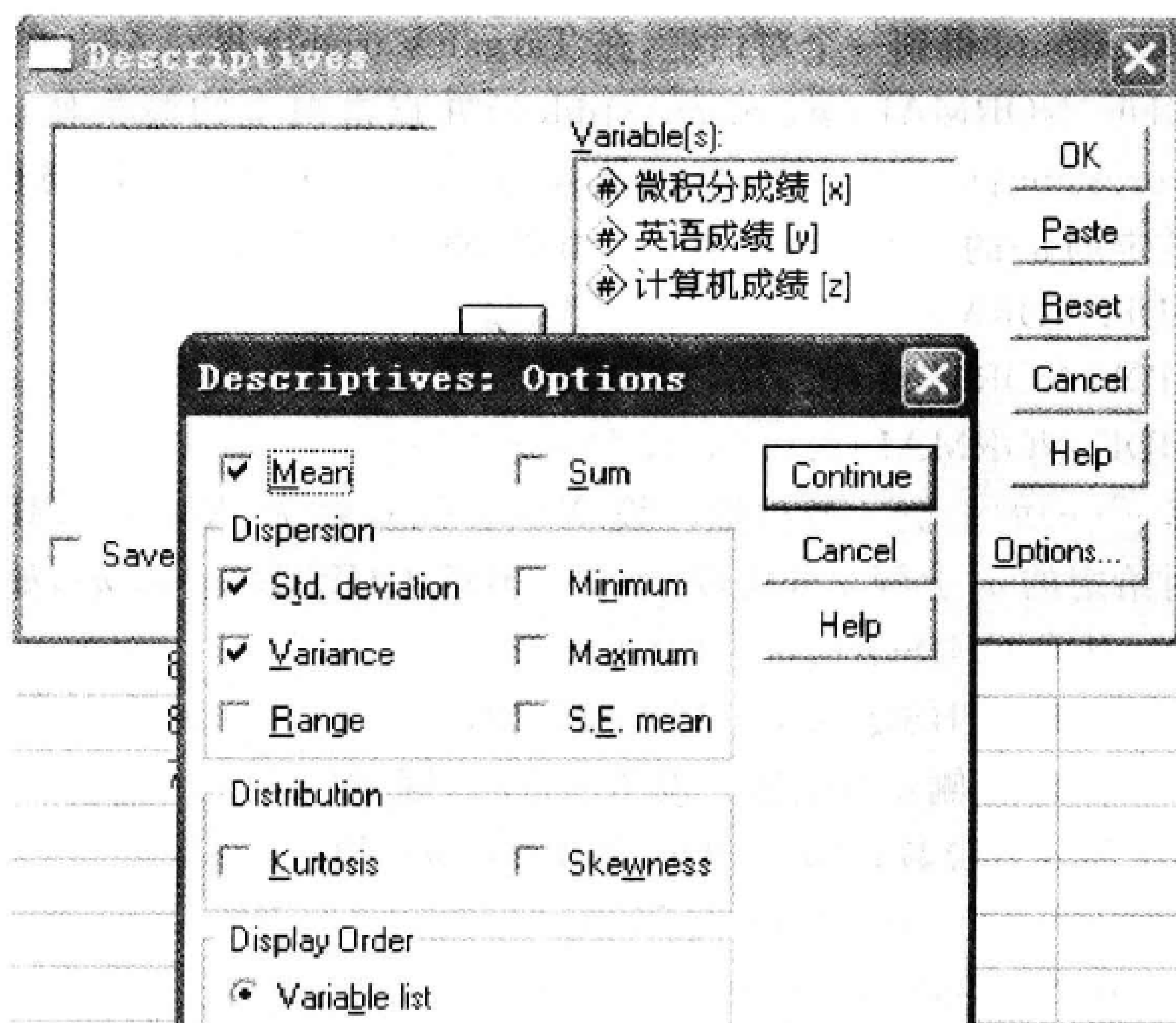


图 5.1.2

点击 Continue→OK，即得到输出结果（表 5.1.2）。

表 5.1.2 **Descriptive Statistics**

	N	Mean	Std. Deviation	Variance
微积分成绩	13	80.92	9.087	82.577
英语成绩	13	81.00	8.794	77.333
计算机成绩	13	80.31	9.961	99.231
Valid N (listwise)	13			

由此表易见，13 个学生的微积分平均成绩为 80.92，标准差为 9.087，方差为 82.577，13 个学生英语平均成绩 81.00，标准差 8.794，方差 77.333，13 个学生的计算机平均成绩为 80.31，标准差 9.961，方差 99.231。

(2) 利用 SPSS 计算四大分布的分位数

① 计算标准正态分布的上侧 α 分位数。因为 $\Phi(z_\alpha)=1-\alpha$，即 CDFNORM $(z_\alpha)=1-\alpha$，在 SPSS 数据编辑窗口(需要随便打开一个数据文件)，依次点击 Tansform→Compute 得到一个对话框，在 Targat Variable 里写入 z，在右侧 Functions：中选 IDF. NORMAL(p, mean, stddev) 并点击向上的箭头使该函数放入 Numeric Expression：中，在三个问号处填入($1-\alpha$, 0, 1)，点击 OK 便得到标准正态分布(对给定的 α)的上侧 α 分位数 z_α = IDF. NORMAL，如：

$z_{0.05}$ = IDF. NORMAL(0.95, 0, 1) = 1.64

$z_{0.025}$ = IDF. NORMAL(0.975, 0, 1) = 1.96

$z_{0.975}$ = IDF. NORMAL(0.025, 0, 1) = −1.96

② 计算 $\chi^2(n)$ 的上侧 α 分位数。设 $X\sim\chi^2(n)$，则 $P(X\leqslant x)$ = IDF. CHISQ (x, n)，而对给定的 α，上侧 α 分位数 $\chi^2_\alpha(n)$ = IDF. CHISQ($1-\alpha$, n)，如：

$\chi^2_{0.95}(n)$ = IDF. CHISQ(0.05, 12) = 5.23

$\chi^2_{0.05}(n)$ = IDF. CHISQ(0.95, 12) = 21.03，

③ 计算 $t(n)$ 的上侧 α 分位数。设 $X\sim t(n)$，则 $P(X\leqslant x)$ = IDF. T(x, n)，而对给定的 α，上侧 α 分位数 $t_\alpha(n)$ = IDF. T($1-\alpha$, n)，如：

$t_{0.05}(n)$ = IDF. T(0.95, 16) = 1.7459

$t_{0.95}(n)$ = IDF. T(0.05, 16) = −1.7459

④ 计算 $F(n_1, n_2)$ 的上侧 α 分位数。设 $X\sim F(n_1, n_2)$，则 $P(X\leqslant x)$ = IDF. F(x, n_1, n_2)，而对给定的 α，上侧 α 分位数 $F_\alpha(n_1, n_2)$ = IDF. F($1-\alpha$, n_1, n_2)，如：

$F_{0.10}(n_1, n_2)$ = IDF. F(0.90, 12, 13) = 2.10

$F_{0.95}(n_1, n_2)$ = IDF. F(0.05, 12, 13) = 0.38

 **四大分布函数值演示实验**

例 5.1.6　在总体 $X\sim N(12,4)$ 中抽取样本容量为 5 的样本 X_1,X_2,X_3,X_4,X_5，求下列概率：① $P(|\overline{X}-12|<1)$；② $P(\max(X_1,\cdots,X_5)>15)$；③ $P(\min(X_1,\cdots,X_5)>10)$

解　① 因为 $\overline{X}\sim N(12,\frac{4}{5})$，所以 $\frac{\overline{X}-12}{\sqrt{\frac{4}{5}}}\sim N(0,1)$，于是

$$P(|\overline{X}-12|<1)=P\left(\left|\frac{\overline{X}-12}{\sqrt{\frac{4}{5}}}\right|<\frac{1}{\sqrt{\frac{4}{5}}}\right)$$
$$=2\Phi(1.118)-1=2\text{CDFNORM}(1.118)-1$$
$$=2\times 0.8682-1=0.7364$$

② $$P(\max(X_1,\cdots,X_5)>15)=1-P(\max(X_1,\cdots,X_5)\leqslant 15)$$
$$=1-P(X_1\leqslant 15,X_2\leqslant 15,\cdots,X_5\leqslant 15)$$
$$=1-P(X_1\leqslant 15)P(X_2\leqslant 15)\cdots P(X_5\leqslant 15)$$
$$=1-[P(X\leqslant 15)]^5=1-[\Phi(1.5)]^5$$
$$=1-[\text{CDFNORM}(1.5)]^5$$
$$=1-(0.93319)^5=1-0.7077=0.2923$$

③ $$P(\min(X_1,\cdots,X_5)>10)=P(X_1>10)P(X_2>10)\cdots P(X_5>10)$$
$$=[P(X>10)]^5=[1-P(X\leqslant 10)]^5$$
$$=[1-\Phi(-1)]^5$$
$$=[\Phi(1)]^5=[\text{CDFNORM}(1)]^5$$
$$=(0.8413)^5=0.4215$$

例 5.1.7　设 $X_1,X_2,\cdots,X_{25}$ 是取自总体 $X\sim N(3,100)$ 的样本，求概率 $P(0<\overline{X}<6,57.7<S^2<151.734)$

解　由抽样分布基本定理可得

$$P(0<\overline{X}<6,57.7<S^2<151.73)=P(0<\overline{X}<6)P(57.7<S^2<151.73)$$
$$=P(-1.5<\frac{\overline{X}-3}{2}<1.5)P(13.848<\frac{24}{100}S^2<36.4152)$$
$$=[2\text{CDFNORM}(1.5)-1][\text{CDF.CHISQ}(36.4152,24)-\text{CDF.CHISQ}(13.848,24)]$$
$$=[2\times 0.9332-1][0.9500-0.0500]=0.7798$$

练习 5.1

1. 设总体 $X\sim f(x)=\begin{cases}|x|, & |x|<1\\ 0, & |x|\geqslant 1\end{cases}$ $X_1, X_2, \cdots, X_{50}$ 为取自该总体的样本，求(1)样本均值的数学期望和方差；(2)样本方差的数学期望；(3)样本均值的绝对值大于 0.02 的概率。

2. 设总体 $X\sim N(\mu, \sigma^2)$，假如要以 0.9606 的概率保证偏差 $|\overline{X}-\mu|<0.1$，问：当 $\sigma^2=0.25$ 时，样本容量应取多大？

3. 从一个正态总体 $X\sim N(\mu, \sigma^2)$ 中抽取容量为 10 的样本，且 $P(|\overline{X}-\mu|>4)=0.02$，求：$\sigma$。

4. 设在总体 $X\sim N(\mu, \sigma^2)$ 中抽取一个容量为 16 的样本，μ，σ^2 均未知，求 $P\left(\dfrac{S^2}{\sigma^2}\leqslant 1.664\right)$。

5. 设总体 $X\sim N(\mu, 16)$，$X_1, X_2, \cdots, X_{10}$ 为取自该总体的样本，已知 $P(S^2>a)=0.1$，求：常数 a。

6. 设总体 $X\sim N(\mu, \sigma^2)$，$X_1, X_2, \cdots, X_n$ 为取自该总体的样本，求：(1) $P((\overline{X}-\mu)^2\leqslant\dfrac{\sigma^2}{n})$；(2)当样本容量很大时，$P((\overline{X}-\mu)^2\leqslant\dfrac{2S^2}{n})$；(3)当样本容量等于 6 时，$P((\overline{X}-\mu)^2\leqslant\dfrac{2S^2}{3})$。

7. 设 $X_1, X_2, \cdots, X_{10}$ 为取自总体 $X\sim N(0, 0.09)$ 的样本，求：$P(\sum\limits_{i=1}^{10}X_i^2>1.44)$。

8. 设 $X_1, X_2, \cdots, X_n$ 是取自总体 $X\sim N(\mu, \sigma^2)$ 的样本，$\overline{X}$ 为样本均值，又记 $S_1^2=\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\overline{X})^2$，$S_2^2=\dfrac{1}{n}\sum\limits_{i=1}^{n}(X_i-\overline{X})^2$，$S_3^2=\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\mu)^2$，$S_4^2=\dfrac{1}{n}\sum\limits_{i=1}^{n}(X_i-\mu)^2$，则服从分布 $t(n-1)$ 的随机变量 $T=$ ______。

(A) $\dfrac{\overline{X}-\mu}{\dfrac{S_1}{\sqrt{n-1}}}$ (B) $\dfrac{\overline{X}-\mu}{\dfrac{S_2}{\sqrt{n-1}}}$ (C) $\dfrac{\overline{X}-\mu}{\dfrac{S_3}{\sqrt{n-1}}}$ (D) $\dfrac{\overline{X}-\mu}{\dfrac{S_4}{\sqrt{n-1}}}$

9. 若 $T\sim t(n)$，则 T^2 服从什么分布？

10. 设 $X_1, X_2, \cdots, X_9$ 为取自总体 $X\sim N(0, 4)$ 的样本，求常数 a, b, c 使得 $Q=a(X_1+X_2)^2+b(X_3+X_4+X_5)^2+c(X_6+X_7+X_8+X_9)^2$ 服从 χ^2 分布，并求

其自由度。

11. 设有 k 个正态总体 $X \sim N(\mu_i, \sigma^2)$，从第 i 个总体中抽取容量为 n_i 的样本 $X_{i1}, X_{i2}, \cdots, X_{in_i}$，且各组样本间相互独立，记 $\overline{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$，$i = 1, 2, \cdots, k$，$n = n_1 + n_2 + \cdots + n_k$，求：$W = \frac{1}{\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2$ 的分布。

12. 设随机变量 X, Y 相互独立且都服从标准正态分布，而 $X_1, X_2, \cdots, X_9$ 和 $Y_1, Y_2, \cdots, Y_9$ 分别是取自总体 X, Y 的相互独立的简单随机样本，求统计量 $Z = \frac{X_1 + X_2 + \cdots + X_9}{\sqrt{Y_1^2 + Y_2^2 + \cdots + Y_9^2}}$ 的分布，并指明参数。

13. 设 $X_1, X_2, \cdots, X_9$ 是取自正态总体 $X \sim N(\mu, \sigma^2)$ 的样本，且 $Y_1 = \frac{1}{6}(X_1 + X_2 + \cdots + X_6)$，$Y_2 = \frac{1}{3}(X_7 + X_8 + X_9)$，$S^2 = \frac{1}{2}\sum_{i=7}^{9}(X_i - Y_2)^2$，求证：$Z = \frac{\sqrt{2}(Y_1 - Y_2)}{S} \sim t(2)$。

14. 设总体 $X \sim N(\mu, \sigma^2)$，从中取出样本 $X_1, X_2, \cdots, X_n, X_{n+1}$，记 $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$，$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$，求证：$\sqrt{\frac{n}{n+1}}\,\frac{X_{n+1} - \overline{X}_n}{S_n} \sim t(n-1)$。

15. 设总体 $X \sim N(0, 4)$，而 $X_1, X_2, \cdots, X_{15}$ 为取自该总体的样本，则随机变量 $Y = \frac{X_1^2 + X_2^2 + \cdots + X_{10}^2}{2(X_{11}^2 + X_{12}^2 + \cdots + X_{15}^2)}$ 服从__________分布，参数为__________。

16. 设总体 $X \sim N(0, 1)$，$X_1, X_2, \cdots, X_n$ 为取自该总体的样本，求：$V = \left(\frac{n}{5} - 1\right)\frac{\sum_{i=1}^{5} X_i^2}{\sum_{i=6}^{n} X_i^2}\ (n > 5)$ 的分布。

5.2　期望与方差的点估计

1. 参数估计的基本思想

正如前面所说，数理统计的基本任务是依据取自总体的样本对总体进行推断。实际中，有时知道了总体的分布类型，分布便由几个与总体有关的未知的数字所决定。要掌握总体的分布，依据样本对这些未知参数做尽可能准确的推断就显得非常重要。

我们把未知的数字叫做**未知参数(unknown parameter)**，对其进行推断有两个问题需要解决：一是求其近似值，称其为未知参数的**点估计值（量）(point estimate)**；二是求其近似范围，称之为未知参数的**区间估计(interval estimate)**。

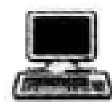
根据待估未知参数的统计意义，由样本构造恰当的统计量，利用样本值计算统计量的值，此统计量的值称为待估未知参数的估计值，因该估计值表现为实轴上的一个点（数），故我们称这种方法为参数的点估计法. 有时并不要求对未知参数做定值估计，而要求估计出未知参数的一个所在范围，并指出该范围包含未知参数的概率，该范围表现为实轴上的一个区间，所以这种方法称为区间估计法。最常见的未知参数是总体的期望和方差，本节介绍其点估计法。

2. 点估计的评选标准

设待估未知参数为 θ，其点估计为 $\hat{\theta}=f(X_1, X_2, \cdots, X_n)$。易见，抽样前 $\hat{\theta}=f(X_1, X_2, \cdots, X_n)$ 为随机变量（估计量），抽样后为一个实数（估计值），对不同的样本值，θ 的估计值 $\hat{\theta}=f(X_1, X_2, \cdots, X_n)$ 也不同。我们自然希望估计值 $\hat{\theta}=f(X_1, X_2, \cdots, X_n)$ 尽可能准确地表达未知参数 θ 的真值，这就产生了对估计值（量）的评选问题。评选估计量 $\hat{\theta}=f(X_1, X_2, \cdots, X_n)$ 的标准一般有以下三个方面。

(1) 无偏性(unbiasedness)

不同的样本值可得不同的估计值，随着抽样的不同，估计值也不同. 由样本所得的估计值可以不等于未知参数的真值，但它们应在真值附近，有时小，有时相等，有时大，很自然应当要求它们的平均值恰好等于真值，即 $E(\hat{\theta})=\theta$。若 θ 的点估计 $\hat{\theta}$ 满足这一要求，我们便称 $\hat{\theta}$ 为 θ 的一个**无偏估计（量）(unbiased estimate)**，即用 $\hat{\theta}$ 做 θ 的近似值没有系统偏差。

 **无偏性演示实验**

例 5.2.1 设 $X_1, X_2, \cdots, X_n$ 是取自总体 X 的样本，总体期望 $E(X)=\mu$ 未知，$a_1, a_2, \cdots, a_n$ 为常数，且 $a_1+a_2+\cdots+a_n=1$，求证：$\sum_{i=1}^{n} a_i X_i$ 为 $E(X)=\mu$ 的一个无偏估计。

证 因为 $E(\sum_{i=1}^{n} a_i X_i)=\sum_{i=1}^{n} a_i E(X_i)=(\sum_{i=1}^{n} a_i)E(X)=\mu$，

所以 $\sum_{i=1}^{n} a_i X_i$ 为 $E(X)=\mu$ 的一个无偏估计。

特别，取 $a_1 = a_2 = \cdots = a_n = \dfrac{1}{n}$，可知样本均值 $\overline{X}$ 为总体均值 μ 的一个无偏估计。

从本例易知两点：一是 μ 的无偏估计不唯一，因为 $a_1, a_2, \cdots, a_n$ 的选择方法有许多种；二是当总体方差 $D(X) = \sigma^2$ 存在时，在 $a_1, a_2, \cdots, a_n$ 的所有选择中，样本均值的方差最小，即 $D(\overline{X}) \leqslant D(\sum\limits_{i=1}^{n} a_i X_i)$，此因

$$
\begin{aligned}
1 = (a_1 + a_2 + \cdots + a_n)^2 &= a_1^2 + \cdots + a_n^2 + 2a_1a_2 + \cdots + 2a_1a_n + 2a_2a_3 + \cdots + 2a_{n-1}a_n \\
&\leqslant a_1^2 + \cdots + a_n^2 + (a_1^2 + a_2^2) + \cdots + (a_{n-1}^2 + a_n^2) \\
&= n(a_1^2 + \cdots + a_n^2)
\end{aligned}
$$

从而可得：$a_1^2 + \cdots + a_n^2 \geqslant \dfrac{1}{n}$。

例 5.2.2　设总体 X 的期望为 μ，方差为 σ^2，$X_1, X_2, \cdots, X_n$ 为取自总体 X 的样本，则样本方差 S^2 为 σ^2 的一个无偏估计，而样本二阶中心矩 B_2 不是 σ^2 的无偏估计。

证　因为
$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$$

利用公式 $E(X^2) = D(X) + (E(X))^2$ 可得

$$
\begin{aligned}
E(\sum_{i=1}^{n}(X_i - \overline{X})^2) &= \sum_{i=1}^{n} E(X_i^2) - nE(\overline{X}^2) \\
&= \sum_{i=1}^{n}(\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2) = (n-1)\sigma^2
\end{aligned}
$$

故 $E(S^2) = \sigma^2$，即样本方差 S^2 为 σ^2 的一个无偏估计。

而 $E(B_2) = \dfrac{n-1}{n}\sigma^2 \neq \sigma^2 (n > 1)$，所以样本二阶中心矩 B_2 不是 σ^2 的无偏估计。

无偏性是点估计的基本要求，它保证 $\hat{\theta}$ 对 θ 的估计只有随机误差，而没有系统误差。

(2) 有效性(effectiveness)

因为 θ 的无偏估计是不唯一的，那么自然就产生一个问题，就是在 θ 的所有无偏估计中哪一个更“好”一点呢？必须明白的是，这里“好”的意思是：$\hat{\theta}$ 的取值更靠近 θ 或更集中在 θ 的附近。由此一个自然的评选标准就是方差。

设 $\hat{\theta}_1$ 与 $\hat{\theta}_2$ 均为 θ 的无偏估计，若 $D(\hat{\theta}_1) \leqslant D(\hat{\theta}_2)$，则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 更有效。进一步，如果固定样本容量 n，在 θ 的一切无偏估计量中，$\hat{\theta}$ 的方差达到最小，则称 $\hat{\theta}$ 为 θ 的**有效估计量(efficient estimator)**。易见，X_1 与 $\overline{X}$ 均为 $E(X) = \mu$ 的无偏估

计，但 $D(X_1)=D(X)=\sigma^2$，$D(\overline{X})=\frac{\sigma^2}{n}$，故当 $n>1$ 时，$\overline{X}$ 比 X_1 有效。

例 5.2.3 设 X_1，X_2，…，X_n 是取自总体 X 的样本，取统计量 $\hat{\mu}=\sum_{i=1}^{n}a_iX_i$ 作为 $E(X)=\mu$ 的估计，其中 $a_1+a_2+\cdots+a_n=1$，求常数 a_1，a_2，…，a_n 使得 $\hat{\mu}=\sum_{i=1}^{n}a_iX_i$ 最有效。

解 在例 5.2.1 中已证，$\hat{\mu}=\sum_{i=1}^{n}a_iX_i$ 为 $E(X)=\mu$ 的无偏估计，并且 $D(\hat{\mu})=(a_1^2+a_2^2+\cdots+a_n^2)D(X)$，要使 $a_1^2+a_2^2+\cdots+a_n^2$ 在满足 $a_1+a_2+\cdots+a_n=1$ 的条件下达到最小值，利用拉格朗日乘数法，容易求得 $a_1=a_2=\cdots=a_n=\frac{1}{n}$，即样本均值 $\overline{X}$ 为总体均值 μ 的这种形式的估计量中最有效的估计量。

在实际应用中，常用 Rao-Cramer 不等式来证明 $\hat{\theta}$ 是有效估计量。可以证明：无偏估计量 $\hat{\theta}$ 的方差 $D(\hat{\theta})$ 永远不会小于正数 $D_0(\theta)$，即

$$D(\hat{\theta})\geqslant D_0(\theta)=\frac{1}{nE\left[\frac{\partial}{\partial\theta}\ln f(X,\theta)\right]^2}>0$$

上式称为 Rao-Cramer 不等式。$D_0(\theta)$ 称为方差的下界，当 $D(\hat{\theta})=D_0(\theta)$ 时，$\hat{\theta}$ 即为 θ 的有效估计量。

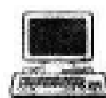
例 5.2.4 设总体 $X\sim P(\lambda)$，X_1，X_2，…，X_n 为取自该总体的样本，试证：$\hat{\lambda}=\overline{X}$ 为 λ 的有效估计量。

证 首先 $E(\hat{\lambda})=E(\overline{X})=\lambda$，知 $\hat{\lambda}=\overline{X}$ 为 λ 的无偏估计量，又因 $D(\hat{\lambda})=D(\overline{X})=\frac{\lambda}{n}$，$f(x,\lambda)=\frac{\lambda^x}{x!}e^{-\lambda}$，$\ln f(x,\lambda)=-\lambda+x\ln\lambda-\ln(x!)$，所以就有 $E\left[\frac{\partial}{\partial\lambda}\ln f(X,\lambda)\right]^2=E\left[\frac{1}{\lambda^2}(X-\lambda)^2\right]=\frac{1}{\lambda^2}D(X)=\frac{1}{\lambda}$，故由 Rao-Cramer 不等式知 $D_0(\theta)=\frac{1}{nE\left[\frac{\partial}{\partial\theta}\ln f(X,\theta)\right]^2}=D(\hat{\lambda})=D(\overline{X})=\frac{\lambda}{n}$，即为 $\hat{\lambda}=\overline{X}$ 为 λ 的有效估计量。

(3) 相合性(consistency)

无论是无偏的或有效的估计量 $\hat{\theta}$，对给定的样本容量 n，一般都不会有 $\hat{\theta}=\theta$。但是当 n 无限增大时，$\hat{\theta}$ 可以无限趋于 θ。即对任意给定的正数 ε，总有 $\lim_{n\to\infty}P(|\hat{\theta}-\theta|<\varepsilon)=1$，或 $\lim_{n\to\infty}P(|\hat{\theta}-\theta|\geqslant\varepsilon)=0$。这自然应该成为评选估计量 $\hat{\theta}$ 的第三个标准。

我们称满足此条件的 $\hat{\theta}$ 为 θ 的相合估计(量)。由切比雪夫不等式易见，当 $\lim\limits_{n\to\infty}E|\hat{\theta}-\theta|^r=0$ 对某 $r>0$ 成立时，$\hat{\theta}$ 为 θ 的**相合估计(量)(consistent estimator)**。

 **相合性演示实验**

例 5.2.5 设 $X_1, X_2, \cdots, X_n$ 为取自总体 X 的样本，$E(X)=\mu$, $D(X)=\sigma^2$，则样本均值 $\overline{X}$ 是总体均值 $E(X)=\mu$ 的相合估计量。

证 利用切比地雪夫不等式得

$$P(|\overline{X}-\mu|\geqslant\varepsilon)\leqslant\frac{D(\overline{X})}{\varepsilon^2}=\frac{\sigma^2}{\varepsilon^2 n}$$

所以 $\lim\limits_{n\to\infty}P(|\overline{X}-\mu|\geqslant\varepsilon)=0$

故样本均值 $\overline{X}$ 是总体均值 $E(X)=\mu$ 的相合估计量。

在现代统计学的研究中，均方误差

$$E(\hat{\theta}-\theta)^2=D(\hat{\theta})+[E(\hat{\theta})-\theta]^2$$

是评选估计量 $\hat{\theta}$ 的最常用也是最全面的标准。由此，近几十年来，提出了许多有用的并已付诸实施的均方误差小的有偏估计，如岭估计、主成分估计、根方估计、偏最小二乘估计等等，可以说这些研究确实成为统计学研究领域一道亮丽的风景线。

3. 数学期望和方差的点估计

数学期望和方差是总体 X 的两个最重要的数字特征。数学期望表示 X 取值的平均水平，方差表示 X 取值相对于平均水平的偏离程度。因此，利用取自总体 X 的样本 $X_1, X_2, \cdots, X_n$ 的均值 $\overline{X}$ 作为总体数学期望 μ 的估计量，用样本方差 S^2 作为总体方差 σ^2 的估计量是很自然的。利用估计量的评选标准也充分证明了这样做的合理性。下面我们再将这种方法规范出来。

例 5.2.6 设总体 $X\sim U(a, b)$，$X_1, X_2, \cdots, X_n$ 为取自该总体的样本，求 a, b 的估计量。

解 因为 $E(X)=\dfrac{a+b}{2}$, $D(X)=\dfrac{(b-a)^2}{12}$,

令
$$E(X)=\overline{X},\ D(X)=S^2,$$

可得方程组
$$\frac{a+b}{2}=\overline{X},\ \frac{(b-a)^2}{12}=S^2$$

解之得 $\hat{a}=\overline{X}-\sqrt{3}S$, $\hat{b}=\overline{X}+\sqrt{3}S$ 即为所求。

练习 5.2

1. 设总体 $X\sim U(\theta, 2\theta)$，其中 $\theta>0$ 是未知参数，又 X_1，X_2，…，X_n 为取自该总体的样本，$\overline{X}$ 为样本均值. 证明：$\hat{\theta}=\dfrac{2}{3}\overline{X}$ 是参数 θ 的无偏估计。

2. 设 $\hat{\theta}$ 是参数 θ 的无偏估计量，$D(\hat{\theta})>0$，证明：$\hat{\theta}^2$ 不是 θ^2 的无偏估计量。

3. 设 X_1，X_2，X_3 是取自总体 X 的样本，试证下列统计量都是总体均值 μ 的无偏估计量，并指出哪一个最有效？

(1) $\hat{\mu}_1=\dfrac{1}{2}X_1+\dfrac{1}{3}X_2+\dfrac{1}{6}X_3$； (2) $\hat{\mu}_2=\dfrac{1}{3}X_1+\dfrac{1}{3}X_2+\dfrac{1}{3}X_3$；

(3) $\hat{\mu}_3=\dfrac{1}{6}X_1+\dfrac{1}{6}X_2+\dfrac{2}{3}X_3$；

4. 设总体 $X\sim U(0, \theta)$，现从该总体中抽取容量为 10 的样本，样本值为：

0.5，1.3，0.6，1.7，2.2，1.2，0.8，1.5，2.0，1.6

试对参数 θ 进行点估计。

5. 从一批电子元件中抽取 8 个进行寿命测试，得到如下数据(单位：小时)：

1050，1100，1130，1040，1250，1300，1200，1080

试对这批元件的平均寿命以及寿命分布的标准差进行点估计。

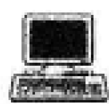
6. 从均值为 μ，方差为 σ^2 的正态总体中分别抽取容量为 n_1 和 n_2 的两组独立样本 $\overline{X}_1$，$\overline{X}_2$ 分别为两组样本的样本均值。试证：对任何常数 a，$b(a+b=1)$，$Y=a\overline{X}_1+b\overline{X}_2$ 都是 μ 的无偏估计，并确定 a，b 的值使 $Y=a\overline{X}_1+b\overline{X}_2$ 在此形式的估计量中最有效。

5.3 期望、方差的区间估计及 SPSS 实现

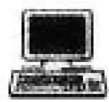
如前所述，点估计是用一个点(即一个实数)去估计未知参数，通俗地讲，就是找到了未知参数的一个近似值。顾名思义，区间估计就是用一个区间去估计未知参数。即在实际问题中，通常需要知道未知参数的一个近似范围。比如，估计一个人的年龄在 20～25 岁之间，估计某人某一次旅游所需费用在 1000～1500 元之间，等等。区间估计是一种很常用的估计形式，其好处是把可能的误差用醒目的形式标出来了。你估计旅游费用需 1000 元，我相信多少会有误差，误差多少？单从你提出的 1000 这个数字还给不出什么信息。你若估计旅游费用在 800～

1200 元之间，则人们会相信你在作出估计时，已把可能出现的误差考虑到了，多少给人们以更大的信任感。

用数学语言来表述，所谓区间估计，即依据取自总体的样本 X_1，X_2，…，X_n，找两个统计量 $\hat{\theta}_1$，$\hat{\theta}_2$，使得 $P(\hat{\theta}_1<\theta<\hat{\theta}_2)=1-\alpha$，其中 $0<\alpha<1$，这时，区间 $(\hat{\theta}_1,\hat{\theta}_2)$ 称为未知参数 θ 的**置信区间(confidence interval)**，$\hat{\theta}_1$ 称为**置信下限(lower confidence limit)**，$\hat{\theta}_2$ 称为**置信上限(upper confidence limit)**，$1-\alpha$ 称为**置信度(degree of confidence)**或**置信水平(confidence level)**或**置信系数(confidence coefficient)**。易见，α 为区间 $(\hat{\theta}_1,\hat{\theta}_2)$ 不含 θ 的概率，即对未知参数估计失准的概率。

 **置信区间几何解释演示实验**

容易明白，α 越小，即 $1-\alpha$ 越大，说明区间 $(\hat{\theta}_1,\hat{\theta}_2)$ 的可靠性越好，另一方面，区间长度 $\hat{\theta}_2-\hat{\theta}_1$ 越小，说明区间 $(\hat{\theta}_1,\hat{\theta}_2)$ 的精确性越高。比如，估计一个人的年龄在某一区间内，如 (30,35) 内，我们要求估计尽量可靠，即该人的年龄有很大把握在这个区间内。同时也要求区间不能太长，比如，估计一人的年龄在区间(10,90)内，当然可靠了，但精确性太差，用处不大。但是，可靠性、精确性这两个要求是相互矛盾的。区间估计理论和方法的基本问题，就是在已有的样本资源的限制下，怎样找出更好的估计方法，以尽量提高可靠性和精确性，但终归有一定的限度。波兰裔美国统计学家 J・奈曼在 20 世纪 30 年代提出并为现代所广泛接受的原则是：先保证可靠度，在保证足够可靠性的基础上尽量提高精确性。

 **区间估计性质演示实验**

按 J・奈曼的这个原则，就是在保证给定的置信系数 $1-\alpha$ 之下，去寻找有优良精度的区间估计，而这个“优良”，也可以有种种原则，目前已有一些结果. 鉴于本书范围，我们所能做的，就是从直观出发，如何去构造看来是合理的区间估计。

1. 单正态总体数学期望的区间估计

设总体 $X\sim N(\mu,\sigma^2)$，X_1，X_2，…，X_n 为取自该总体的样本。

(1) σ^2 已知，μ 待估

由于要推断总体数学期望 μ，自然想到了它的无偏估计样本均值 $\bar{X}$，依抽样

分布基本定理，$\overline{X}\sim N(\mu,\frac{\sigma^2}{n})$，标准化后可得：

$$Z=\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}=\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\sim N(0,1)$$

注意到 Z 有以下三个特点：一是样本的函数；二是含且仅含待估未知参数 μ；三是其分布与待估未知参数 μ 无关。为方便计，有人喜欢称具有这三个特点的变量为“枢轴变量”，它一般是从待估未知参数的一个良好的点估计出发来构造的。

对给定的 α，即置信度 $1-\alpha$，因为 $P(|Z|<z_{\frac{\alpha}{2}})=1-\alpha$，解不等式 $|Z|<z_{\frac{\alpha}{2}}$，即得 $\overline{X}-\frac{\alpha}{\sqrt{n}}z_{\frac{\alpha}{2}}<\mu<\overline{X}+\frac{\sigma}{\sqrt{n}}z_{\frac{\alpha}{2}}$，这说明在置信度 $1-\alpha$ 下，得到未知参数 μ 的置信区间为

$$(\overline{X}-\frac{\sigma}{\sqrt{n}}z_{\frac{\alpha}{2}},\ \overline{X}+\frac{\sigma}{\sqrt{n}}z_{\frac{\alpha}{2}}) \tag{5.3.1}$$

(2) σ^2 未知，μ 待估

由于 σ^2 未知，所以此时 Z 不能再用，实际中，一个自然的想法是用 σ^2 的无偏估计样本方差 S^2 替代之，这时便得到

$$T=\frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}}=\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}\sim t(n-1)$$

这样得到的随机变量 T 也恰好具有与 Z 类似的三个特点，也为“枢轴变量”，于是对给定的 α，由于 $P(|T|<t_{\frac{\alpha}{2}}(n-1))=1-\alpha$，解不等式 $|T|<t_{\frac{\alpha}{2}}(n-1)$，可得 $\overline{X}-\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)<\mu<\overline{X}+\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)$，于是，当总体方差 σ^2 未知的情况下总体期望 μ(未知参数)的置信度为 $1-\alpha$ 的置信区间为

$$(\overline{X}-\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)),\ \overline{X}+\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)) \tag{5.3.2}$$

下面我们介绍利用 Excel 求置信区间的方法。Excel 求置信区间使用 CONFIDENCE函数，其语法格式如下

$$\text{CONFIDENCE}(\alpha,\sigma,n)=\frac{\sigma}{\sqrt{n}}z_{\frac{\alpha}{2}} \tag{5.3.3}$$

置信下限为：

$$\bar{X} - \mathrm{CONFIDENCE}(\alpha, \sigma, n) \tag{5.3.4}$$

置信上限为：

$$\bar{X} + \mathrm{CONFIDENCE}(\alpha, \sigma, n) \tag{5.3.5}$$

例 5.3.1　设正态总体的方差为 1，根据取自该总体的容量为 100 的样本计算得到样本均值为 5，求总体均值的置信度为 0.95 的置信区间。

解　由题设条件知，应选用公式

$$\left(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}},\ \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}\right)$$

又 $\sigma=1, n=100, \bar{x}=5, \alpha=0.05, z_{0.025}=$ IDF. NORMAL(0.975, 0, 1)$=1.96$，所以，$\frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}} = \frac{1}{10} \times 1.96 = 0.196$，$\bar{x}-0.196=4.804$，$\bar{x}+0.196=5.196$，从而所求置信区间为(4.804，5.196)。

例 5.3.2　某种零件的重量服从正态分布．现从中抽取容量为 16 的样本，其观测到的重量(单位：千克)分别为 4.8，4.7，5.0，5.2，4.7，4.9，5.0，5.0，4.6，4.7，5.0，5.1，4.7，4.5，4.9，4.9。需要估计零件平均重量，求平均重量的区间估计，置信系数是 0.95。

解　零件的平均重量的点估计是 $\bar{X}$，因零件重量的方差未知，我们必须用式(5.3.2)来求总体均值的区间估计。

下面我们介绍应用 SPSS 求置信区间的方法。首先，建立数据文件(表 5.3.1，见 SPSS 数据文件例 5.3.2)

表 5.3.1　随机抽取的 16 个零件的重量

	x
1	4.8
2	4.7
3	5.0
4	5.2
5	4.7
6	4.9
7	5.0
8	5.0
9	4.6
10	4.7
11	5.0
12	5.1
13	4.7
14	4.5
15	4.9
16	4.9

依次点击 Analyze→Compare Means→One-Sample T Test…,如图5.3.1所示。

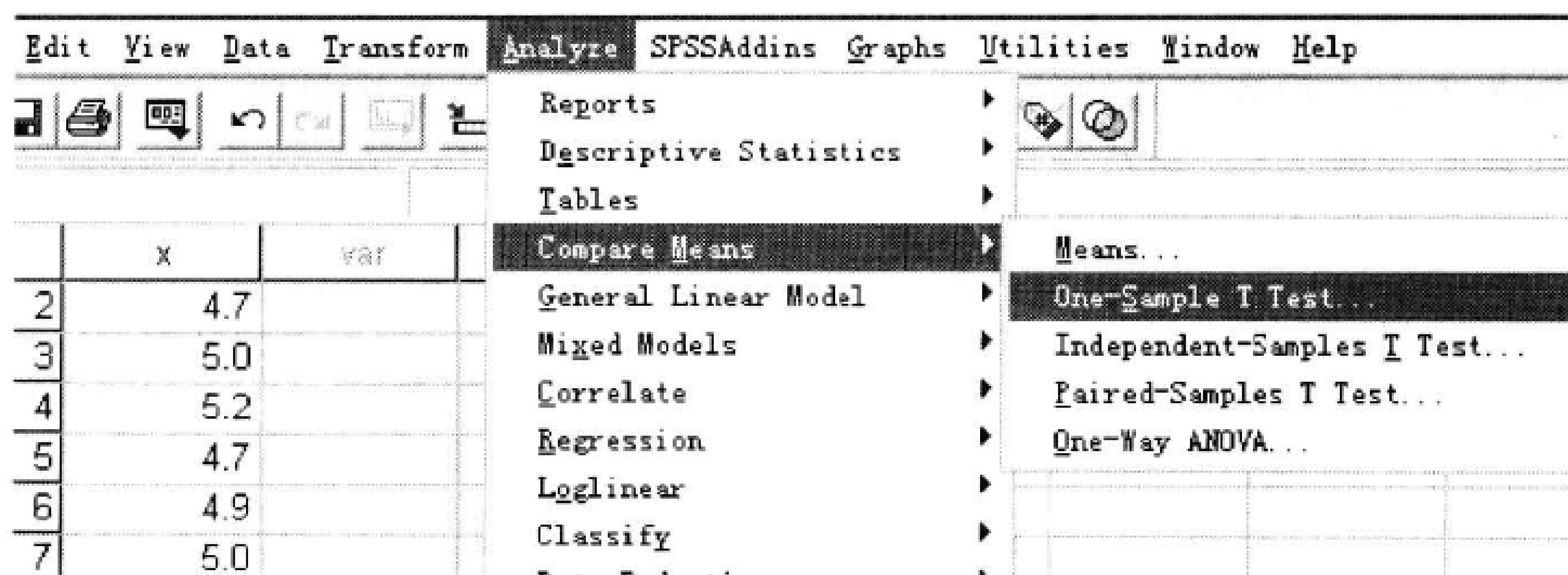


图 5.3.1

在弹出的对话框中,将左侧源变量栏中的 x 选入 Test Variable(s):框中,可以点击 Options... 按钮来设置置信系数,如图 5.3.2。

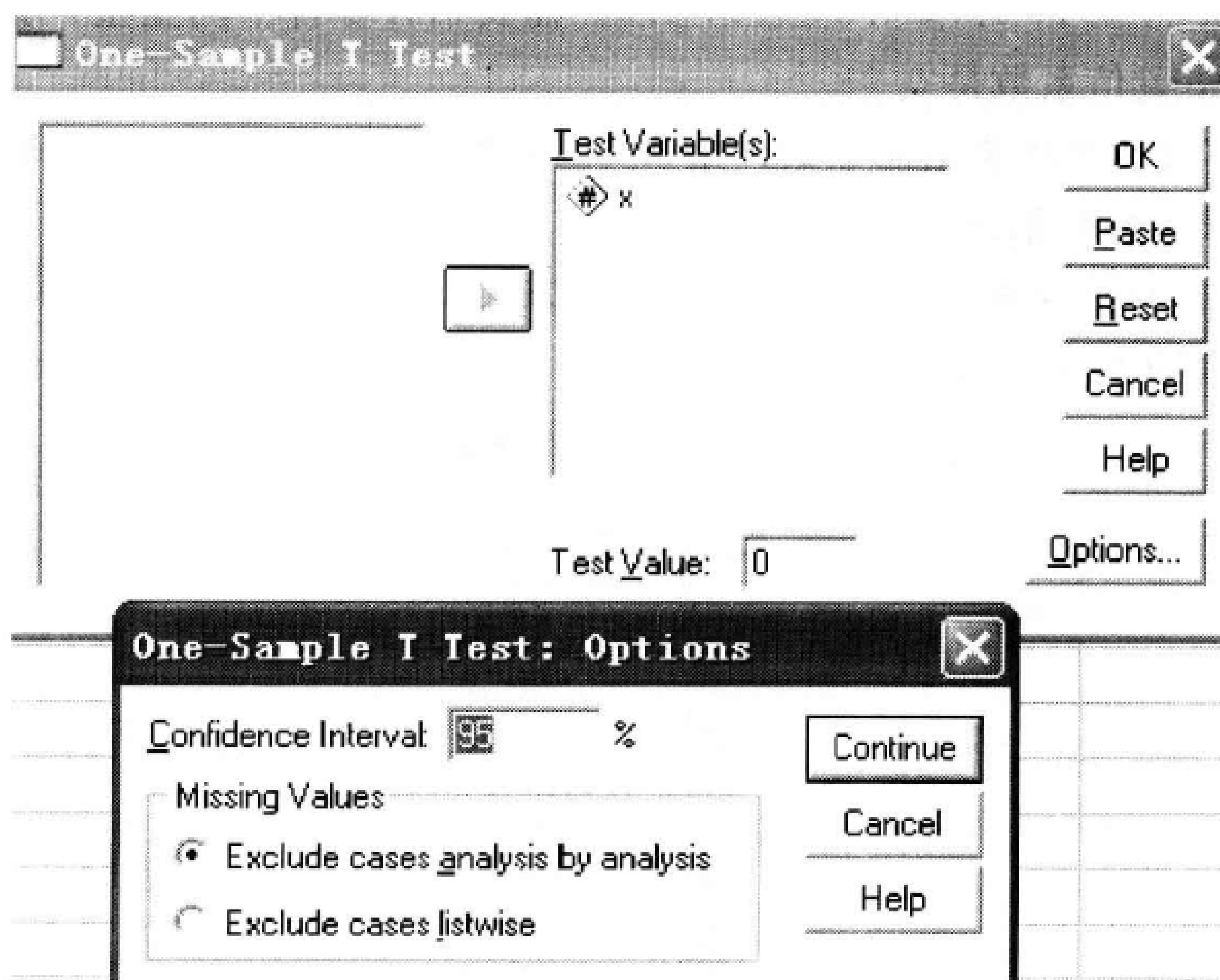


图 5.3.2

点击 OK,即得到输出结果(见表 5.3.2 和表 5.3.3)。

表 5.3.2　One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
X	16	4.856	.1931	.0483

表 5.3.3　One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
X	100.590	15	.000	4.856	4.753	4.959

从表 5.3.3 可见，平均重量的置信系数是 0.95 的区间估计为(4.753，4.959)。

2. 单正态总体方差的区间估计

(1) 总体均值 μ 已知

这时总体方差 σ^2 的无偏估计为 $\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$，且 $Q=\frac{n\hat{\sigma}^2}{\sigma^2}\sim\chi^2(n)$，易见 Q 是“枢轴变量”。对给定的 α，由于 $P(\chi^2_{1-\frac{\alpha}{2}}(n)<Q<\chi^2_{\frac{\alpha}{2}}(n))=1-\alpha$，所以，解不等式 $\chi^2_{1-\frac{\alpha}{2}}(n)<Q<\chi^2_{\frac{\alpha}{2}}(n)$，可得 σ^2 的置信度为 $1-\alpha$ 的置信区间是

$$\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\frac{\alpha}{2}}(n)},\ \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n)}\right) \tag{5.3.6}$$

(2) 总体均值 μ 未知

这时总体方差 σ^2 的无偏估计为样本方差 S^2，且 $Q=\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$，完全类似于上段可得 σ^2 的置信度为 $1-\alpha$ 的置信区间是

$$\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)},\ \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right) \tag{5.3.7}$$

或

$$\left(\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\chi^2_{\frac{\alpha}{2}}(n-1)},\ \frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right) \tag{5.3.7'}$$

例 5.3.3　投资的回收利用率常常用来衡量投资的风险. 随机地调查了 26

个年回收利润率(%)，标准差 $S=15(\%)$。设回收利润率为正态分布，求它的方差的区间估计(置信系数为 0.95)。

解　依题意应用公式(5.3.7)来求解,公式为

$$\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)},\ \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right)$$

因为 $n-1=25$，$S^2=15^2=225$，$\alpha=0.05$,利用 SPSS 计算分位数 $\chi^2_{0.025}(25)=40.65$，$\chi^2_{0.975}(25)=13.12$，

$$\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}=\frac{25\times 225}{40.65}=138.376,\ \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}=\frac{25\times 225}{13.12}=428.735,$$

所以方差的置信区间为:(138.376，428.735)。

3. 两个正态总体均值差的区间估计

设样本 $X_1, X_2, \cdots, X_{n_1}$ 取自总体 $X\sim N(\mu_1, \sigma_1^2)$，样本 $Y_1, Y_2, \cdots, Y_{n_2}$ 取自总体 $Y\sim N(\mu_2, \sigma_2^2)$，且两组样本相互独立，$\overline{X}$，$S_1^2$，$\overline{Y}$，$S_2^2$ 分别表示两组样本的均值和方差。

(1) 当 σ_1^2 和 σ_2^2 均已知时，$\mu_1-\mu_2$ 的区间估计

由于 $\overline{X}\sim N(\mu_1, \frac{\sigma_1^2}{n_1})$，$\overline{Y}\sim N(\mu_2, \frac{\sigma_2^2}{n_2})$，且 $\overline{X}$，$\overline{Y}$ 相互独立，所以 $\overline{X}\sim\overline{Y}\sim N(\mu_1-\mu_2, \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2})$,于是可知

$$Z=\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0, 1)$$

易见，此变量为“枢轴变量”，对给定的概率 α，由 $P(|Z|<z_{\frac{\alpha}{2}})=1-\alpha$，解不等式 $|Z|<z_{\frac{\alpha}{2}}$，可得 $\mu_1-\mu_2$ 的置信度为 $1-\alpha$ 的置信区间是

$$\left(\overline{X}-\overline{Y}-z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\ \overline{X}-\overline{Y}+z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right) \tag{5.3.8}$$

思考：$\mu_2-\mu_1$ 的置信度为 $1-\alpha$ 的置信区间呢?

(2) 当 σ_1^2 和 σ_2^2 均未知，但 $\sigma_1^2=\sigma_2^2=\sigma^2$ 时，$\mu_1-\mu_2$ 的区间估计

我们记 $S_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$，则易证 S_p^2 为 σ^2 的一个无偏估计，且 $T=\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2)$，这个变量显然也是一个“枢轴变量”,利用

它，并与前一段类似的做法可得 $\mu_1-\mu_2$ 的置信度为 $1-\alpha$ 的置信区间是

$$\left(\bar{X}-\bar{Y}-S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}t_{\frac{\alpha}{2}}(n_1+n_2-2),\ \bar{X}-\bar{Y}+S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}t_{\frac{\alpha}{2}}(n_1+n_2-2)\right) \tag{5.3.9}$$

(3) 当 σ_1^2 和 σ_2^2 均未知，但 $n_1=n_2=n$，$\mu_1-\mu_2$ 的区间估计

此时，令 $Z_i=X_i-Y_i$，则 $Z_i\sim N(\mu_1-\mu_2, \sigma_1^2+\sigma_2^2)$，将 $Z_1, Z_2, \cdots, Z_n$ 视为取自总体 $Z\sim N(\mu_1-\mu_2, \sigma_1^2+\sigma_2^2)$的样本，由单正态总体当方差未知时总体均值的区间估计方法可得 $\mu_1-\mu_2$ 的置信度为 $1-\alpha$ 的置信区间是

$$\left(\bar{Z}-\frac{S_z}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1),\ \bar{Z}+\frac{S_z}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)\right) \tag{5.3.10}$$

其中 $\bar{Z}=\bar{X}-\bar{Y}$，$S_z^2=\frac{1}{n-1}\sum_{i=1}^{n}(Z_i-\bar{Z})^2$。

(4) 当 σ_1^2 和 σ_2^2 均未知，但 n_1，n_2 均很大(一般 n_1，n_2 均大于 50)，$\mu_1-\mu_2$ 的区间估计

这是所谓的大样本情形，可用

$$\left(\bar{X}-\bar{Y}-z_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}},\ \bar{X}-\bar{Y}+z_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right) \tag{5.3.11}$$

作为 $\mu_1-\mu_2$ 的置信度为 $1-\alpha$ 的近似置信区间。

(5) 当 σ_1^2 和 σ_2^2 均未知，$\mu_1-\mu_2$ 的区间估计

在这种情况下，求 $\mu_1-\mu_2$ 的区间估计是统计学中的一个著名的问题，叫做贝伦斯-费希尔问题。这个问题经过许多著名的统计学家研究过，得不出简单确切的解法，但提出了一些近似解法(所谓“近似解法”，其含义就是所求出的区间估计的置信系数不一定严格地等于预定的 $1-\alpha$，而只是近似地等于它)。下面我们仅介绍这些解法中的一个。

分别以两组样本的样本方差 S_1^2，S_2^2 去估计两总体的方差 σ_1^2，σ_2^2 而得到 $\lambda^2=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$的估计为 $\hat{\lambda}^2=\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}$，得到变量 $T=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\hat{\lambda}}$。这样，严格地说，该变量已不再服从 t 分布，但与具有适当自由度 k 的 t 分布 $t(k)$很接近。k 由公式

$$k=\frac{\hat{\lambda}}{\frac{S_1^4}{(n_1-1)n_1^2}+\frac{S_2^4}{(n_1-1)n_2^2}} \tag{5.3.12}$$

决定(k 一般不为整数，可以取与它最接近的整数代替之)，近似地取 $T\sim t(k)$ ，

应用与情况(2)类似的步骤，得到 $\mu_1-\mu_2$ 的区间估计为

$$(\overline{X}-\overline{Y}-\hat{\lambda}t_{\frac{\alpha}{2}}(k),\ \overline{X}-\overline{Y}+\hat{\lambda}t_{\frac{\alpha}{2}}(k)) \tag{5.3.13}$$

置信系数近似地等于 $1-\alpha$。

例 5.3.4 某工厂利用两条自动化流水线听装番茄酱，分别从两条流水线上抽取随机样本：$X_1, X_2, \cdots, X_{12}$ 和 $Y_1, Y_2, \cdots, Y_{17}$，计算出 $\overline{X}=10.6$(克)，$\overline{Y}=9.5$(克)，$S_1^2=2.4$，$S_2^2=4.7$。假设这两条流水线上听装番茄酱的重量都服从正态分布，其总体均值分别为 μ_1，μ_2，且有相同的总体方差。试求总体均值差 $\mu_1-\mu_2$ 的区间估计，置信系数为 0.95。

解 依题意，利用公式(5.3.9)求解。

$$\left(\overline{X}-\overline{Y}-S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}t_{\frac{\alpha}{2}}(n_1+n_2-2),\ \overline{X}-\overline{Y}+S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}t_{\frac{\alpha}{2}}(n_1+n_2-2)\right)$$

因为 $n_1=12$，$n_2=17$，$\alpha=0.05$，$S_1^2=2.4$，$S_2^2=4.7$，$\bar{x}=10.6$，$\bar{y}=9.5$，又计算：

$$S_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}=\frac{11\times2.4+16\times4.7}{27}=3.7630,\ S_p=1.940$$

$$t_{\frac{\alpha}{2}}(n_1+n_2-2)=t_{0.025}(27)=2.05$$

$$\overline{X}-\overline{Y}-S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}t_{\frac{\alpha}{2}}(n_1+n_2-2)=10.6-9.5-1.940\times\sqrt{\frac{1}{12}+\frac{1}{17}}\times2.05=-0.3995$$

$$\overline{X}-\overline{Y}+S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}t_{\frac{\alpha}{2}}(n_1+n_2-2)=10.6-9.5+1.940\times\sqrt{\frac{1}{12}+\frac{1}{17}}\times2.05=2.5995$$

所以，总体均值差 $\mu_1-\mu_2$ 的区间估计为(-0.3995 , 2.5995)。

4. 两正态总体方差比的区间估计

由于 S_1^2，S_2^2 分别为 σ_1^2，σ_2^2 的无偏估计，又因为 $\frac{(n_1-1)S_1^2}{\sigma_1^2}\sim\chi^2(n_1-1)$，$\frac{(n_2-1)S_2^2}{\sigma_2^2}\sim\chi^2(n_2-1)$，且二者相互独立，所以

$$F=\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}\sim F(n_1-1,\ n_2-1)$$

易见，此变量为“枢轴变量”。对给定的 α，因为

$$P(F_{1-\frac{\alpha}{2}}(n_1-1,\ n_2-1)<F<F_{\frac{\alpha}{2}}(n_1-1,\ n_2-1))=1-\alpha$$

解不等式，并注意到 $F_{1-\frac{\alpha}{2}}(n_1-1,\ n_2-1)=[F_{\frac{\alpha}{2}}(n_2-1,\ n_1-1)]^{-1}$ 可得总体方差比 σ_1^2/σ_2^2 的置信度为 $1-\alpha$ 的置信区间是

$$\left(\left[F_{\frac{\alpha}{2}}(n_1-1, n_2-1)\right]^{-1}\frac{S_1^2}{S_2^2}, F_{\frac{\alpha}{2}}(n_2-1, n_1-1)\frac{S_1^2}{S_2^2}\right) \tag{5.3.14}$$

或

$$\left(F_{1-\frac{\alpha}{2}}(n_2-1, n_1-1)\frac{S_1^2}{S_2^2}, F_{\frac{\alpha}{2}}(n_2-1, n_1-1)\frac{S_1^2}{S_2^2}\right) \tag{5.3.14'}$$

或

$$\left(\left[F_{\frac{\alpha}{2}}(n_1-1, n_2-1)\right]^{-1}\frac{S_1^2}{S_2^2}, \left[F_{1-\frac{\alpha}{2}}(n_1-1, n_2-1)\right]^{-1}\frac{S_1^2}{S_2^2}\right) \tag{5.3.14''}$$

思考：σ_2^2/σ_1^2 的置信度为 $1-\alpha$ 的置信区间？

例 5.3.5　为了比较用两种不同方法生产的某种产品的寿命而进行一项试验。试验中抽选了由方法一生产的 16 个产品组成一随机样本，其方差为 1200 小时；又抽选了由方法二生产的 21 个产品组成另一随机样本，得出的方差为 800 小时。试以 95％的可靠性估计两总体方差之比的置信区间。

解　设方法一生产的产品的寿命为 $X \sim N(\mu_1, \sigma_1^2)$，方法二生产的产品的寿命 $Y \sim N(\mu_2, \sigma_2^2)$，现在要求 σ_1^2/σ_2^2 的置信度为 95％的置信区间。

由题意我们利用公式(5.3.14′)来求解。

$$\left(F_{1-\frac{\alpha}{2}}(n_2-1, n_1-1)\frac{S_1^2}{S_2^2}, F_{\frac{\alpha}{2}}(n_2-1, n_1-1)\frac{S_1^2}{S_2^2}\right)$$

因为 $S_1^2=1200, S_2^2=800$，$\alpha=0.05$，$n_1=16$，$n_2=21$，又易计算

$$F_{\frac{\alpha}{2}}(n_2-1, n_1-1)=F_{0.025}(20, 15)=2.76$$

$$F_{1-\frac{\alpha}{2}}(n_2-1, n_1-1)=F_{0.975}(20, 15)=0.39$$

$$F_{1-\frac{\alpha}{2}}(n_2-1, n_1-1)\frac{S_1^2}{S_2^2}=0.39\times\frac{1200}{800}=0.58$$

$$F_{\frac{\alpha}{2}}(n_2-1, n_1-1)\frac{S_1^2}{S_2^2}=2.76\times\frac{1200}{800}=4.14$$

所以，两总体方差之比 σ_1^2/σ_2^2 的置信度为 95％的置信区间为(0.58，4.14)。

由上述方法求得的总体均值差或总体方差比的置信区间，我们在实际中通常有下列结论：

① 若 $\mu_1-\mu_2$ 的置信区间的下限大于零，则可认为 $\mu_1>\mu_2$；若 $\mu_1-\mu_2$ 的置信区间的上限小于零，则可认为 $\mu_1<\mu_2$；若 $\mu_1-\mu_2$ 的置信区间包含零，则可认为 $\mu_1=\mu_2$。

② 若 σ_1^2/σ_2^2 的置信区间的下限大于 1，则可认为 $\sigma_1^2>\sigma_2^2$；若 σ_1^2/σ_2^2 的置信区间的上限小于 1，则可认为 $\sigma_1^2<\sigma_2^2$；若 σ_1^2/σ_2^2 的置信区间包含 1，则可认为 $\sigma_1^2=\sigma_2^2$。

练习 5.3

1. 由取自正态总体 $X\sim N(\mu, 0.9^2)$，容量为 9 的样本，若得到样本均值为 $\overline{X}=5$,则未知参数 μ 的置信度为 0.95 的置信区间是__________。

2. 已知某种材料的抗压强度 $X\sim N(\mu, \sigma^2)$，现随机地抽取 10 个试件进行抗压试验，测得数据如下：482，493，457，471，510，446，435，418，394，469。

① 求平均抗压强度 μ 的点估计值；

② 求平均抗压强度 μ 的 95% 的置信区间；

③ 若已知 $\sigma=30$，求平均抗压强度 μ 的 95%的置信区间；

④ 求 σ^2 的点估计值；

⑤ 求 σ^2 的 95%的置信区间；

⑥ 求 σ 的点估计值；

⑦ 求 σ 的 95%的置信区间。

3. 设总体 $X\sim N(\mu, \sigma^2)$，σ^2 已知，问样本容量 n 取多大时才能保证 μ 的 95%的置信区间的长度不大于 k。

4. 假设 0.50 ，1.25 ，0.80 ，2.00 是取自总体 X 的样本值，已知 $Y=\ln X$ 服从正态分布 $X\sim N(\mu, 1)$。

① 求 X 的数学期望 $E(X)$（并记 $E(X)=b$）；

② 求 μ 的置信度为 95%的置信区间；

③ 利用上述结果求 b 的置信度为 95%的置信区间。

5. 设从总体 $X\sim N(\mu_1, 64)$和总体 $X\sim N(\mu_2, 36)$中分别抽取容量为 $n_1=75$，$n_2=50$ 的独立样本，可计算得 $\bar{x}=82$，$\bar{y}=76$，求 $\mu_1-\mu_2$ 的 96%的置信区间。

6. 假设人体身高服从正态分布，今抽测甲、乙两地区 18 ～ 25 岁女青年身高得数据如下：甲地区抽取 10 名，样本均值 1.64 米，样本标准差 0.2 米；乙地区抽取 10 名，样本均值 1.62 米，样本标准差 0.4 米。

求：①两正态总体方差比的 99%的置信区间；②两正态总体均值差的 99%的置信区间。

5.4 点估计法

在第 5.2 节中，我们寻找总体未知参数的点估计都是根据未知参数的特征（或统计意义）去寻找具有同类特征（或统计意义）的统计量来实现的。本节我们介绍两种一般化的求点估计的方法。

1. 矩估计法

所谓**矩估计法**(**moment method of estimate**)，就是将总体的各阶原点矩用相应阶的样本原点矩替代，布列方程或方程组所得到的解，作为总体未知参数的点估计的方法。下面我们通过例子来加以说明。

例 5.4.1　设总体 $X \sim U(0, \theta)$，X_1，X_2，…，X_n 为取自该总体的样本，求未知参数 θ 的矩估计量。

解　因为 $E(X)=\frac{\theta}{2}$，所以由 $\frac{\theta}{2}=\overline{X}$，可解得 $\theta=2\overline{X}$，故未知参数 θ 的矩估计量为 $\hat{\theta}=2\overline{X}$。

例 5.4.2　设总体的概率密度函数为

$$f(x)=\begin{cases}\frac{6x(\theta-x)}{\theta^3}, & 0<x<\theta \\ 0, & 其它\end{cases}$$

X_1，X_2，…，X_n 为取自该总体的样本。

求：①未知参数的矩估计量 $\hat{\theta}$；②$D(\hat{\theta})$。

解　① 容易计算出 $E(X)=\frac{\theta}{2}$，令 $\frac{\theta}{2}=\overline{X}$，解得 $\theta=2\overline{X}$，

所以，未知参数 θ 的矩估计量为 $\hat{\theta}=2\overline{X}$。

② 因为 $D(\hat{\theta})=4D(\overline{X})=4\times\frac{D(X)}{n}$，而 $E(X)=\frac{\theta}{2}$，$E(X^2)=\frac{6\theta^2}{20}$，

所以
$$D(X)=E(X^2)-(E(X))^2=\frac{\theta^2}{20},$$

于是
$$D(\hat{\theta})=\frac{\theta^2}{5n}。$$

2. 最大似然估计法(maximum-likelihood method of estimate)

设总体 X 的密度函数为 $f(x, \theta)$，X_1，X_2，…，X_n 为取自该总体的样本。由第 3 章的概率论知识可知，随机向量$(X_1, X_2, \cdots, X_n)$的联合密度函数为

$$L=L(x_1, x_2, \cdots, x_n; \theta)=f(x_1, \theta)f(x_2, \theta)\cdots f(x_n, \theta),$$

当未知参数 θ 固定，而看作是 x_1，x_2，…，x_n 的函数时，L 是一个概率密度函数或概率函数，它可以这样理解：若 $L(y_1, y_2, \cdots, y_n; \theta)>L(x_1, x_2, \cdots, x_n; \theta)$，则在对总体进行观测时，出现点$(y_1, y_2, \cdots, y_n)$的可能性要比出现点$(x_1, x_2, \cdots, x_n)$的可能性大。把这件事反过来可以这样想，当 x_1，x_2，…，x_n 已知(已经观测到$(x_1, x_2, \cdots, x_n)$时)，若 $L(x_1, x_2, \cdots, x_n; \theta_1)>(L(x_1, x_2, \cdots, x_n; \theta_2)$，

则被估计的未知参数 θ 是 θ_1 的可能性要比是 θ_2 的可能性大。当 x_1，x_2，…，x_n 固定，而把 L 看作是 θ 的函数时，称 L 为“**似然函数(likelihood function)**”，这名称的意义可根据上述分析得到。这函数对不同的 θ 的取值反映了在观测结果 x_1，x_2，…，x_n 已知的条件下 θ 的各种值的“似然程度”。注意这里有些像贝叶斯公式中的推理，把观测值看作结果，把参数值看作导致这结果的原因，现已有了结果，要反过来推算各种原因的概率。这里参数 θ 有一定的值(虽然未知)，并非事件或随机变量，无概率可言，于是就改用“似然”这个词。从以上分析就自然导致如下的方法：应该用似然程度最大的那个 θ，即满足条件

$$L(x_1, x_2, \cdots, x_n; \hat{\theta}) = \max_{\theta} L(x_1, x_2, \cdots, x_n; \theta)$$

的 $\hat{\theta}$ 去作为 θ 的估计值。因为在已得样本 x_1，x_2，…，x_n 的条件下，这个 $\hat{\theta}$“看来最像”是真参数值。这个估计 $\hat{\theta}$ 就叫做 θ 的**最大似然估计(maximum-likelihood estimate)**。由于函数 $\ln x$ 的单调增加性，所以使 L 达到最大等价于使 $\ln L$ 达到最大，因此，我们经常使用对数似然函数 $\ln L$ 来求最大似然估计。

例 5.4.3　设总体 X 的概率密度函数为

$$f(x) = \begin{cases} (\theta+1)x^{\theta}, & 0 < x < 1 \\ 0, & \text{其它} \end{cases}$$

其中未知参数 $\theta > -1$，X_1，X_2，…，X_n 为取自该总体的样本。求：θ 的最大似然估计量 $\hat{\theta}$ 及 $E(\mathrm{e}^{-\frac{n}{1+\hat{\theta}}})$。

解　首先写出似然函数为

$$\begin{aligned} L = L(x_1, x_2, \cdots, x_n; \theta) &= f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta) \\ &= \begin{cases} (\theta+1)^n (x_1 x_2 \cdots x_n)^{\theta}, & 0 < x_1, x_2, \cdots, x_n < 1 \\ 0, & \text{其它} \end{cases} \end{aligned}$$

当 $0 < x_1$，x_2，…，$x_n < 1$ 时，$\ln L = n\ln(\theta+1) + \theta\ln(x_1 x_2 \cdots x_n)$，令 $\frac{\mathrm{d}}{\mathrm{d}\theta}\ln L = \frac{n}{\theta+1} + \ln(x_1 x_2 \cdots x_n) = 0$，可解得 $\theta = -\frac{n}{\ln(x_1 x_2 \cdots x_n)} - 1$，所以，$\theta$ 的最大似然估计为：$\hat{\theta} = -\frac{n}{\ln(X_1 X_2 \cdots X_n)} - 1$。

又因为 $\mathrm{e}^{-\frac{n}{1+\hat{\theta}}} = X_1 X_2 \cdots X_n$，

所以 $E(\mathrm{e}^{-\frac{n}{1+\hat{\theta}}}) = E(X_1 X_2 \cdots X_n) = [E(X)]^n = (\frac{\theta+1}{\theta+2})^n$。

其中 $E(X) = \int_0^1 (\theta+1) x^{\theta+1} \mathrm{d}x = \frac{\theta+1}{\theta+2}$。

例 5.4.4　设总体 $X \sim B(n, p)$，求 p 的最大似然估计。

解　设 $X_1, X_2, \cdots, X_n$ 为取自 $X \sim B(n, p)$ 的样本，似然函数为

$$L = L(x_1, x_2, \cdots, x_n; \theta) = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta)$$
$$= C_n^{x_1} p^{x_1} (1-p)^{n-x_1} C_n^{x_2} p^{x_2} (1-p)^{n-x_2} \cdots C_n^{x_n} p^{x_n} (1-p)^{n-x_n}$$

对数似然函数为

$$\ln L = \ln(C_n^{x_1} C_n^{x_2} \cdots C_n^{x_n}) + (\sum_{i=1}^{n} x_i) \ln p + \sum_{i=1}^{n} (n - x_i) \ln(1-p)$$

令

$$\frac{d}{dp} \ln L = \frac{1}{p}(\sum_{i=1}^{n} x_i) + \frac{1}{1-p} \sum_{i=1}^{n} (n - x_i) = 0$$

解得

$$p = \frac{1}{n}\bar{x}$$

故 $\hat{p} = \frac{1}{n}\bar{X}$ 为 p 的最大似然估计。

例 5.4.5　设 $X_1, X_2, \cdots, X_n$ 是从服从均匀分布 $U(0, \theta)$ 的总体中抽取的样本，求 θ 的最大似然估计。

解　因为总体 X 的分布密度函数为

$$f(x) = \begin{cases} \frac{1}{\theta}, & 0 < x < \theta \\ 0, & \text{其它} \end{cases}$$

所以，似然函数为

$$L = L(x_1, x_2, \cdots, x_n; \theta) = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta)$$
$$= \begin{cases} \frac{1}{\theta^n}, & 0 < x_1, x_2, \cdots, x_n < \theta \\ 0, & \text{其它} \end{cases}$$

为使似然函数 L 达到最大，θ 必须尽量地小，但又不能太小以致 L 为 0。注意到 $0 < x_1, x_2, \cdots, x_n < \theta$，即 $0 < \min(x_1, x_2, \cdots, x_n)$，$\max(x_1, x_2, \cdots, x_n) < \theta$ 所以取 $\hat{\theta} = \max(X_1, X_2, \cdots, X_n)$，易见，当 $\theta \geqslant \hat{\theta}$ 时，$L = \frac{1}{\theta^n} > 0$，当 $\theta < \hat{\theta}$ 时，$L = 0$，故唯一使 L 达到最大的 θ 值，即 θ 的最大似然估计为 $\hat{\theta} = \max(X_1, X_2, \cdots, X_n)$。

比较例 5.4.1 与例 5.4.5，可见，同一个总体未知参数的矩估计和最大似然估计可能是不相同的，直观上可知，θ 的最大似然估计 $\hat{\theta} = \max(X_1, X_2, \cdots, X_n)$ 不是无偏估计，而是偏低的，事实上，我们可以计算出 θ 的最大似然估计为 $\hat{\theta} = \max(X_1, X_2, \cdots, X_n)$ 的数学期望为 $\frac{n}{n+1}\theta < \theta$，关于这两个估计的优劣评选读者可进一步研讨。

练习 5.4

1. 已知总体 X 服从瑞利分布，其密度函数为

$$f(x)=\begin{cases}\dfrac{x}{\theta}e^{-\frac{x^2}{2\theta}}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

未知参数 $\theta>0$，X_1，X_2，…，X_n 为取自该总体的样本。求 θ 的矩估计量和最大似然估计量，并问这两个估计量是不是无偏估计量？

2. 设总体 X 的对数函数 $\ln X$ 服从正态分布 $N(\mu, \sigma^2)$，X_1，X_2，…，X_n 为取自总体 X 的样本，求 μ，σ^2 及 $E(X)$的最大似然估计量。

3. 已知总体 X 服从参数为 θ 的泊松分布，其分布律为

$$P(X=k)=\frac{1}{k!}\theta^k e^{-\theta},\ k=0, 1, 2, \cdots, \theta>0$$

X_1，X_2，…，X_n 为取自总体 X 的样本。求① θ 的矩估计量；② θ 的最大似然估计量；③ θ 的无偏估计量。

第 6 章　统计检验

在第 5 章中，我们介绍了统计推断的一类方法 —— 未知参数的统计估值、总体未知参数的点估计与区间估计问题。在有些实际问题中，需要我们知道总体的未知参数有无明显变化，或是否达到既定的要求，或多个总体的某个参数有无明显差异，等等。诸如此类问题，就是我们在本章将要介绍的统计推断的另一类方法 —— 统计检验所要解决的问题。

6.1　统计检验概要

先举两个例子。

例 6.1.1　某工厂生产大批的电子元件，假定元件的寿命服从指数分布 $E(\lambda)$，在实际应用中，我们可以提出许多感兴趣的问题，例如：① 元件的平均寿命如何？② 如果你是消费者，要求平均寿命能够达到某个指定的 k 值（比如 5000 小时），问这批元件可否被不拒绝？

对于第一个问题，因为总体平均为 $\frac{1}{\lambda}$，而 λ 是未知的，我们只好从总体中随机抽取 n 个，测量其寿命为 X_1，X_2，…，X_n 构成样本值，然后依据样本值对 $\frac{1}{\lambda}$ 作出估计，这就是前一章研究过的内容。

对于第二个问题，可能认为（至少对本例而言）：解决了第一个问题也就解决了第二个问题，因为既然用样本均值 $\overline{X}$ 去估计平均寿命，那就看 $\overline{X}$ 是否不小于指定的值 k，若 $\overline{X}\geqslant k$，则不拒绝这批元件，不然就拒绝。应当承认这是一个可以考虑的解法。但还应注意到，样本均值 $\overline{X}$ 估计总体平均寿命有误差，我们必须根据实际需要进行一定的调整。即把不拒绝的准则定为 $\overline{X}\geqslant k_1$，$k_1$ 是某个选定的值，可以小于、等于、大于 k。k_1 定的大些，表示检验更严格，这在对元件质量要求很高且供货渠道较多时可能是适当的。反之，k_1 定的小些，表示检验更宽，这在对元件质量要求不很高或急需这些元件而供货渠道很少时，也可能采取。从统计上说，无论怎样定 k_1，理论上都可能犯两种错误之一：一是元件平均寿命达到需求而被拒绝；一是元件平均寿命达不到需求而被拒绝。这两种错误各有一定的

规律，它们在很大程度上决定了不拒绝准则 $\overline{X} \geqslant k_1$ 中 k_1 的选择。

第二个问题与第一个问题不同，它不是要求对分布中的未知参数作出估计，而是要在两个决定（对本例而言就是不拒绝或拒收该批产品）中选择一个。

例 6.1.2 某食品加工厂用自动包装机包装薯条，定额标准为每袋净重量 0.5 千克，设在正常情况下包装机称得薯条重量服从正态分布 $N(0.5, 0.0001)$，根据长期经验知其标准差不变。为检验包装机工作是否正常，随机抽测 16 袋薯条，其净重（单位：千克）为

0.499，0.514，0.508，0.512，0.498，0.515，0.516，0.511，0.524，0.499，0.499，0.500，0.505，0.497，0.490，0.510，

问：该包装机工作是否正常？

在这个问题中，首先样本是从总体 $N(\mu, 0.0001)$ 中抽取的，包装机工作是否正常，就是看 μ 是否等于 0.5 ，即 $\mu=0.5$ 是否成立？若成立，则认为正常；若不成立，则认为不正常。同样是从 $\mu=0.5$ 与 $\mu\neq0.5$ 两个决定中选择一个。

在实践中，类似的问题还有很多很多。在统计学上，首先是对问题发表“看法”，称之为**“假设(hypothesis)”**；再依据样本，用一定的方法论证这一“假设”是否成立，称之为**“统计检验(statistic test)”**。

一般地，我们称关于总体 X 的各种论断为**统计假设(statistic hypothesis)**，简称假设，用 H 表示。如例 6.1.2 中，可设 $H_0: \mu=0.5$，也可设 $H_1: \mu\neq0.5$，它们是相互对立的假设。通常称 H_0 为基本假设（或原假设，零建设，解消假设），而称 H_1 为 H_0 的对立假设（或备择假设，备选假设）。所有的统计假设都由两部分组成，一个原假设和一个备择假设。所构造出来的这两个假设包含了试验或研究的所有可能结果。一般而言，原假设(null hypothesis)阐述的是“原”条件存在的情形，也即没有新事物出现，原来的理论仍然正确，原来的标准仍然适用，原来的系统没有被打乱。备择假设(alternative hypothesis)则相反，它阐述的是，新理论才是正确的，系统已经混乱以及（或）新事物出现了。再举个例子来说，假定一个面粉加工厂包装的面粉按重量出售；一种特定大小的包装平均质量为 10 kg。假如该厂想检验其包装流程是否正常工作（以面粉包装的重量为依据）。该试验的原假设就是面粉包装的平均重量为 10 kg（工作正常），备择假设是面粉包装的平均重量不是 10 kg（流程出错）。“假设”这个词在此就是一个其正确与否有待通过样本去判断的陈述，不能与数学上常说的“假设……”之类的话相混，后者是一个所讨论的问题中已被承认的前提或条件。而“检验”一词就是对“假设”正确与否的“判断”。

1. 统计检验的基本思想

在例 6.1.2 中，基本假设 $H_0: \mu=\mu_0=0.5$，备择假设 $H_1: \mu\neq\mu_0$。由 μ 的统

计意义，我们自然联想到样本均值 $\overline{X}$，根据样本值，我们可计算出 $\overline{x}$，$\overline{x}$ 与 μ_0 之间存在差异。对这个差异有两个解释，一是这差异由抽样的随机性造成，而 H_0 成立，即包装机工作正常；一是这差异由包装机工作不正常造成，即 H_0 不成立 H_1 成立. 到底哪一个解释比较合理呢？

首先，统计问题中不能要求 $\overline{x}$ 与 μ_0 之间没有差异，从直观的角度分析，如果包装机工作正常，即 H_0 成立，那么 $\overline{x}$ 与 μ_0 之间应该相差不大；反过来，若 $\overline{x}$ 与 μ_0 之间相差很大，我们自然应当怀疑 H_0，认为 H_0 不成立，即包装机工作不正常。我们将"$\overline{x}$ 与 μ_0 的差异"、"应该相差不大"、"相差很大"这样的直观描述定量化。刚才说到"容许有差异"，即 $|\overline{X}-\mu_0|>\lambda$，"应该相差不大"，即"差异不大"的概率要大，或"差异大"的概率要小，即 $P(|\overline{X}-\mu_0|>\lambda)\leqslant\alpha$（很小的正数），这样就考虑了"差异"与 "出现差异"的可能性之间的关系。λ 用来描写"差异"的度，称之为临界值，α 用来描写"差异"的可能性，称之为**显著性水平(significance level)**（或**检验水平(critical value)**）。若在 H_0 成立的条件下，$P(|\overline{X}-\mu_0|>\lambda)\leqslant\alpha$，说明事件$\{|\overline{X}-\mu_0|>\lambda\}$在 H_0 成立下出现的概率很小，称为（条件）小概率事件，而小概率事件在一次试验（或一次抽取）中是几乎不可能出现的，称为实际推断原理（或小概率原理）。若一次抽样，小概率事件$\{|\overline{X}-\mu_0|>\lambda\}$出现了，即 $|\overline{X}-\mu_0|>\lambda$，则认为"差异"超过了一定的度，作出的判断是：$H_0$ 不成立，即要拒绝 H_0。

再回到例 6.1.2 中，对给定的 $\alpha=0.05$，由于 $\overline{X}\sim N(\mu,\frac{0.0001}{16})$，则当 H_0 成立时，就有

$$\frac{\overline{X}-\mu_0}{\sqrt{\frac{0.0001}{16}}}=40(\overline{X}-\mu_0)\sim N(0,1)$$

$$P(|40(\overline{X}-\mu_0)|>z_{\frac{\alpha}{2}})=P(|\overline{X}-\mu_0|>\frac{1}{40}z_{0.025})=0.05=\alpha$$

因此，当 $|\overline{X}-\mu_0|>\frac{1}{40}z_{0.025}$ 成立时，则认为 H_0 不成立，拒绝 H_0，即认为包装机工作不正常；否则，不拒绝 H_0，认为包装机工作正常。

这里再说明一点，在许多文献中，把检验准则 $|\overline{X}-\mu_0|>\frac{\sigma}{\sqrt{n}}z_{\frac{\alpha}{2}}$ 改为考查概率 $p=P(|Z|>\frac{|\overline{X}-\mu_0|}{\frac{\sigma}{\sqrt{n}}})$，其中 $Z\sim N(0,1)$，若 $p<\alpha$，则拒绝 H_0，若 $p\geqslant\alpha$，则不拒绝 H_0。

一般把这两种方法分别称为临界值法和概值法。

2. 统计检验的实施程序

Step1 根据题意明确总体，合理地建立基本假设 H_0 和备择假设 H_1；当 H_1 位于 H_0 的两侧时，被称之为双侧检验；当 H_1 位于 H_0 的一侧时，被称之为单侧检验。

Step2 选择适当的**检验统计量(test statistic)**，要求在 H_0 为真时，统计量的分布是确定和已知的。

Step3 规定检验水平 α，并由基本假设 H_0 和备择假设 H_1 确定一个合理的拒绝 H_0 的条件。

Step4 由样本值计算出检验统计量的值。

Step5 作出判断：若统计量的值满足拒绝条件，则拒绝 H_0，否则不拒绝 H_0。

以上过程可由 SPSS 实现. 在上述程序中，Step1 是前提，Step2 是关键。值得注意的是：做假设检验用的统计量与参数区间估计用的随机变量在形式上是一致的，每一个区间估计法都对应一个假设检验法。

3. 两类错误

当我们根据抽样结果而不拒绝或拒绝一个假设时，这只是表明我们的一种判断。由于样本有随机性，这样作出的判断就可能犯错误。

第一类错误(type one error)：弃真，即 H_0 成立而误认为 H_0 不成立，犯第一类错误是因为拒绝了一个正确的原假设。例如，假定面粉包装流程正常工作，每袋面粉平均重量 10 千克。又假定研究人员随机抽取 100 袋，称出每袋重量后计算样本均值，有时可能碰巧抽取了 100 袋比较极端的(最重的或最轻的)，这样就会得到一个落入拒绝域的均值。结果是拒绝原假设，尽管总体均值确实是 10 千克。在这种情况下，研究人员就犯了第一类错误。在统计假设检验以外的活动中第一类错误的说法也存在。比如，一名经理解雇了一名雇员，因为有证据表明他在公司盗窃，如果他实际上并没有在公司盗窃，那么该经理就犯了第一类错误。再比如，假定一家大型制造企业装配生产线上的一名工人在听到异常声音后决定停止生产线(拒绝原假设)。如果后来发现声音和装配线没有关联，装配线上没有出现任何故障，那么该工人就犯了第一类错误。在法庭上，也有第一类错误的类似例子，那就是一个无辜的人被判入狱。易见，弃真错误出现的概率(最大值)为 α(显著性水平(level of significance))。

第二类错误(type two error)：取伪，即 H_0 不成立而误认为 H_0 成立。当研

究人员没有能拒绝一个错误的原假设时，他就犯了第二类错误。犯第二类错误的概率通常记为 β。与 α 不同的是，β 通常不会在假设检验程序开始时就给出。实际上因为只有当原假设不正确时 β 才存在，所以 β 的计算会因可能存在的备择参数不同而不同。例如，在面粉包装问题中，如果总体均值不是 10 千克，那它是多少呢？可能是 12 千克，也可能是 8 千克。β 的值与每个备择的均值相对应。

α 和 β 有什么关联呢？首先，因为只有当原假设被拒绝时，才存在 α；当不拒绝原假设时，才存在 β。所以在同一次的假设检验中，研究人员不可能同时犯第一类错误和第二类错误。

我们自然希望犯这两类错误的概率越小越好，但样本容量 n 确定之后，犯这两类错误的概率不可能同时被控制，减小其中一个，另一个往往会增大。α 越小，拒绝域就越窄，也即拒绝原假设就更难，接受原假设就更容易。比如对加工装配线，如果管理体制使工人们不能轻易地关闭装配线（减少了第一类错误），那么生产出质量差的产品或装配线发生严重问题的机会就会更大了（增大了第二类错误的概率）。

一种能同时减小两种错误的方法是增加样本容量。如果选取的样本容量变大，那么样本就更能代表总体，这也就意味着研究人员做出正确选择的机会更大。

如果把事物的实际面目称为“本质状态”，研究人员实际做出的决策称为“结论”。而每个备择的“结论”都只包含一类错误，以及作出正确决策的概率。当原假设错误时，拒绝原假设的概率 $1-\beta$ 一般称为“势”。α、β 与势的关系见下表 6.1.1。

表 6.1.1 两类错误

		本质状态	
		原假设正确	原假设错误
结论	不拒绝原假设	正确决策	第二类错误(β)
	拒绝原假设	第一类错误(α)	正确决策（势）

 **两类错误演示实验**

在实践中，通常采用的做法是：控制犯弃真错误的概率不超过某个事先给定的显著性水平 α（很小的正数），而使犯取伪错误的概率尽可能得小。

如果我们仅要求犯弃真错误的概率等于 α，则这类统计检验问题即为显著性检验问题。正因为显著性检验只给出了犯弃真错误的控制标准 α，实际中恰当地选取 H_0 就尤为重要了。一般地，可据以下原则选取 H_0：① 当我们的目的是希望从样本值中取得对某一论断的强有力的支持时，把这一论断的对立面作为基本假

设 H_0；② 尽可能使后果严重的一类错误作为第一类错误；③ 把由历史资料所提供的论断作为基本假设 H_0。当检验结论为拒绝 H_0 时，由于犯弃真错误的概率被控制而显得有说服力和危害较小；当检验结论为不拒绝 H_0 时，因不拒绝的理由是抽样的结果与 H_0 无矛盾，或者说没有找到拒绝 H_0 的理由，故检验结果是没有说服力的，并且犯取伪错误的概率还不受控制，所以通常在不拒绝 H_0 时，为可靠起见，还须进一步进行抽样检验。

6.2 单正态总体的统计检验及 SPSS 实现

设总体 $X \sim N(\mu, \sigma^2)$，X_1，X_2，…，X_n为取自该总体的样本。

1. 期望的检验

(1) 当总体方差 σ^2 已知时

① $H_0: \mu=\mu_0$(已知)，$H_1: \mu \neq \mu_0$。

在 $H_0: \mu=\mu_0$ 成立的条件下，$Z=\dfrac{\overline{X}-\mu_0}{\dfrac{\sigma}{\sqrt{n}}} \sim N(0, 1)$，故对给定的检验水平 α，由 $P(|Z|>z_{\frac{\alpha}{2}})=\alpha$ 可求得：$H_0: \mu=\mu_0$ 的拒绝条件为

$$|Z|>z_{\frac{\alpha}{2}}$$

通常称此检验为 Z 检验。

② $H_0: \mu \leqslant \mu_0$(已知)，$H_1: \mu>\mu_0$。由于在 $H_0: \mu \leqslant \mu_0$ 成立的条件下有 $\dfrac{\overline{X}-\mu_0}{\dfrac{\sigma}{\sqrt{n}}} \leqslant \dfrac{\overline{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}}$，所以对给定的 α 就有 $\left\{\dfrac{\overline{X}-\mu_0}{\dfrac{\sigma}{\sqrt{n}}}>z_\alpha\right\} \subseteq \left\{\dfrac{\overline{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}}>z_\alpha\right\}$，于是 $P\left(\dfrac{\overline{X}-\mu_0}{\dfrac{\sigma}{\sqrt{n}}}>z_\alpha\right) \leqslant P\left(\dfrac{\overline{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}}>z_\alpha\right)=\alpha$，故 $H_0: \mu \leqslant \mu_0$ 的拒绝条件为

$$Z>z_\alpha$$

③ $H_0: \mu \geqslant \mu_0$(已知)，$H_1: \mu<\mu_0$。类似地，$H_0: \mu \geqslant \mu_0$ 的拒绝条件为

$$Z<-z_\alpha$$

(2) 当总体方差 σ^2 未知时

同第 5 章未知参数区间估计一样，只需将检验统计量 Z 中的未知参数总体标准差 σ 替换成样本标准差 S 便可得以上三种情形各对应的基本假设的拒绝条件。现列表 6.2.1 如下。

表 6.2.1

H_0	H_1	总体方差 σ^2 已知 检验统计量 $Z=\dfrac{\overline{X}-\mu_0}{\dfrac{\sigma}{\sqrt{n}}}$ （Z 检验）	总体方差 σ^2 未知 检验统计量 $T=\dfrac{\overline{X}-\mu_0}{\dfrac{S}{\sqrt{n}}}$ （T 检验）
		在显著性水平 α 下的 H_0 的拒绝条件	
$\mu=\mu_0$	$\mu\neq\mu_0$	$\lvert Z\rvert>z_{\frac{\alpha}{2}}$	$\lvert T\rvert>t_{\frac{\alpha}{2}}(n-1)$
$\mu\leqslant\mu_0$	$\mu>\mu_0$	$Z>z_\alpha$	$T>t_\alpha(n-1)$
$\mu\geqslant\mu_0$	$\mu<\mu_0$	$Z<-z_\alpha$	$T<-t_\alpha(n-1)$

例 6.2.1　两厂生产同一产品，其质量指标假定都服从正态分布，标准规格为均值等于 120。现从甲厂抽出 5 件产品测得其指标值为

119.0，120.0，119.2，119.7，119.6

从乙厂也抽取 5 件产品，测得其指标值为

110.5，106.3，122.2，113.8，117.2

要根据这些数据去判断这两厂产品是否符合预定规格 120?（显著性水平 0.05）。

解　设甲厂产品指标服从正态分布 $N(\mu_1, \sigma_1^2)$，乙厂产品指标服从正态分布 $N(\mu_2, \sigma_2^2)$。σ_1^2 和 σ_2^2 均未知。

首先，建立数据文件表 6.2.2（见 SPSS 数据文件例 6.2.1）。

表 6.2.2　两厂抽取的产品质量指标

Edit　View　Data　Transform　Analyze

	x	y	var
1	119.0	110.5	
2	120.0	106.3	
3	119.2	122.2	
4	119.7	113.8	
5	119.6	117.2	
6			

对甲厂，H_0：$\mu_1=\mu_0=120$，H_1：$\mu_1\neq\mu_0=120$ 进行双侧 T 检验。

对乙厂，H'_0：$\mu_2=\mu_0=120$，H'_1：$\mu_2\neq\mu_0=120$ 进行双侧 T 检验。

其次，在 SPSS 数据编辑窗口，依次调用程序如图 6.2.1 所示。

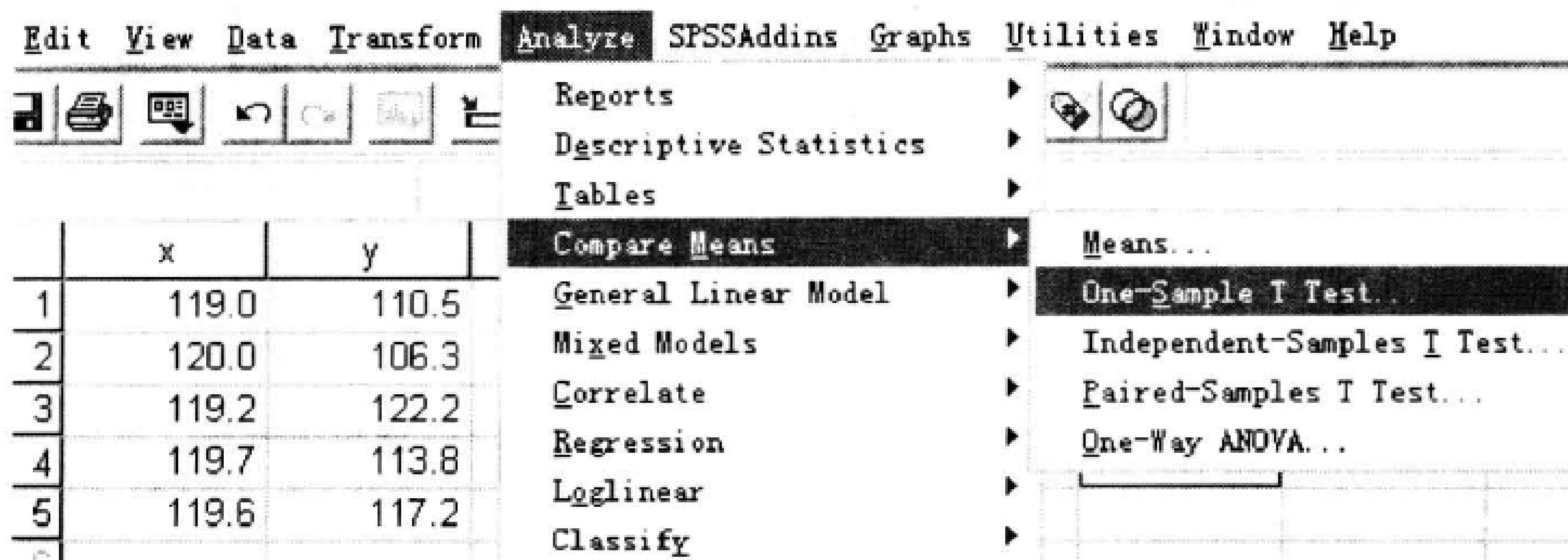


图 6.2.1

得到对话框并按照需要勾选相应的项（如图 6.2.2）。

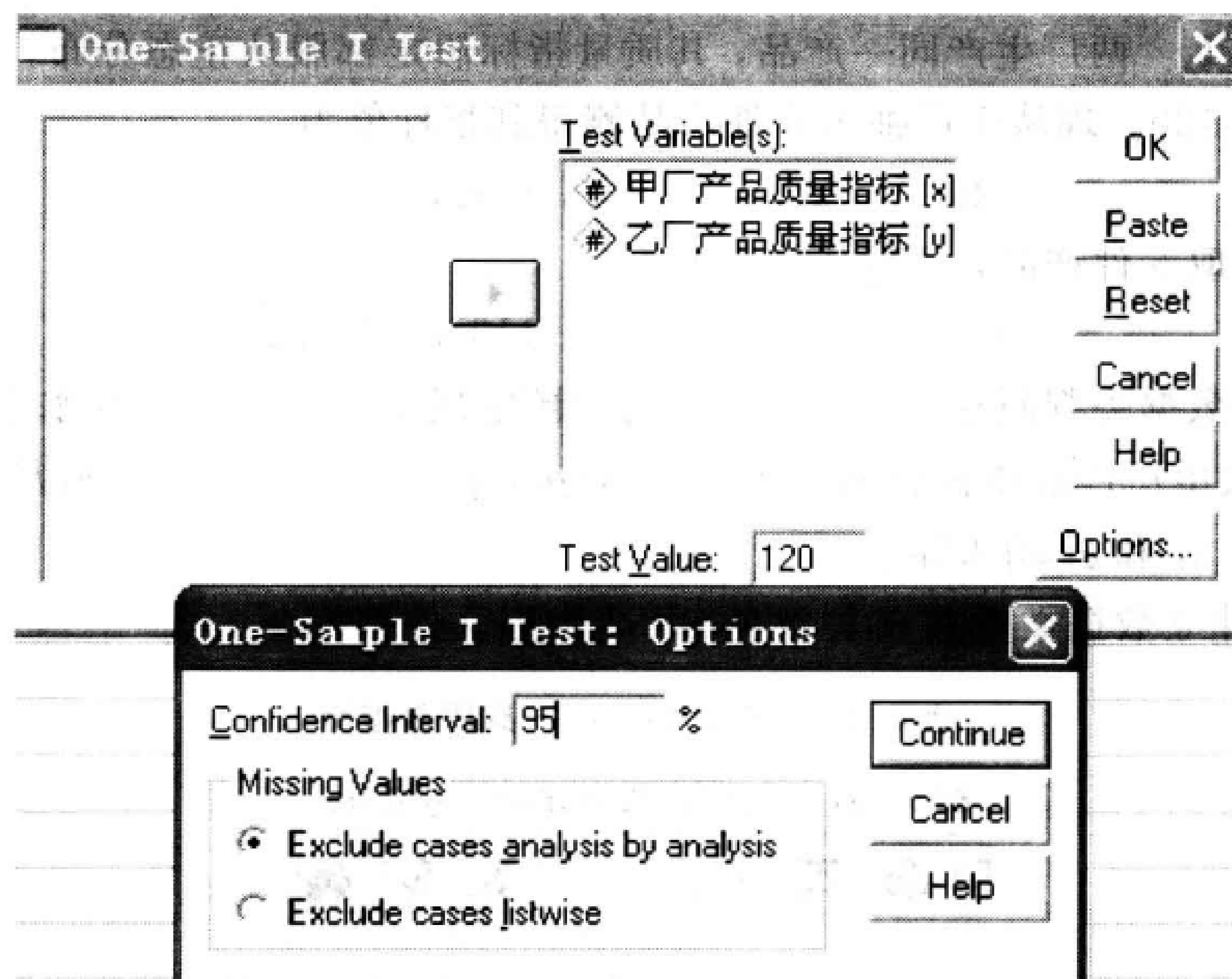


图 6.2.2

再点击 Continue→OK，便得到输出结果如表 6.2.3 和表 6.2.4。

表 6.2.3 One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
甲厂产品质量指标	5	119.500	.4000	.1789
乙厂产品质量指标	5	114.000	6.1045	2.7300

表 6.2.4　**One-Sample Test**

	Test Value = 120					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
甲厂产品质量指标	−2.795	4	.049	−.500	−.997	−.003
乙厂产品质量指标	−2.198	4	.093	−6.000	−13.580	1.580

由于 $t_{\frac{\alpha}{2}}(n-1)=t_{0.025}(4)=2.776$，而对于甲厂 $t=-2.795$，因为 $|-2.795|>2.776$，所以应当拒绝基本假设 H_0：$\mu_1=\mu_0=120$，即甲厂产品与预定规格不符。对于乙厂 $t=-2.198$，因为 $|-2.198|<2.776$，所以结论是不拒绝基本假设 H'_0：$\mu_2=\mu_0=120$，即未发现乙厂产品不符合预定规格的有力证据。

以上检验方法称为"临界值法"，此外，我们还经常用另一种检验方法，称为"概值法"。对于甲厂而言，即计算概值 $p=P(t(n-1)>|t|)=P(t(4)>2.795)=0.049$，因为 $p<0.05=\alpha$，所以也应当拒绝基本假设 H_0：$\mu_1=\mu_0=120$，即认为甲厂产品与预定规格不符。对于乙厂而言，$p=0.093>0.05=\alpha$，所以应当不拒绝基本假设 H'_0：$\mu_2=\mu_0=120$，即未发现乙厂产品不符合预定规格的有力证据。

p 值的确定方法如图 6.2.3 示。

如上结论可能使不少人感到难以接受，因为甲厂 5 件产品都与标准规格相差很少，反倒认为不合格，而乙厂 5 件中除一件外，都比规格值低不少，反倒认为可以通过，这是为什么？

首先，甲厂的 $S=0.4$ 远低于乙厂的 $S=6.10507$，这表明甲厂产品规格比乙厂稳定得多，也正因为甲厂产品规格很齐整（误差很小），所以，与规格值 120 的稍微差别（此处 $\bar{x}=119.5$ 比 120 仅差 0.5）也被检出来了。不能不承认：甲厂产品的平均规格，有很大可能略低于规格值 120，虽只略低些，也是事实，不能委之于随机误差。至于这样一个差别的实际重要性如何，那要另当别论了。此处只讲统计上的显著性——即误差不能用随机误差去解释。统计上显著不一定有现实重要性。

乙厂抽出的 5 件产品有 4 件的指标远低于规格值 120，使我们很有理由怀疑：乙厂产品平均指标达不到规格值 120，但由于乙厂产品质量波动太大，所测数据尚不能很有把握认为其平均指标确与 120 有差距，而非随机性影响所致，就是说，现有数据可能太少了些。所以对乙厂我们首先认为：其产品质量波动太大应当改

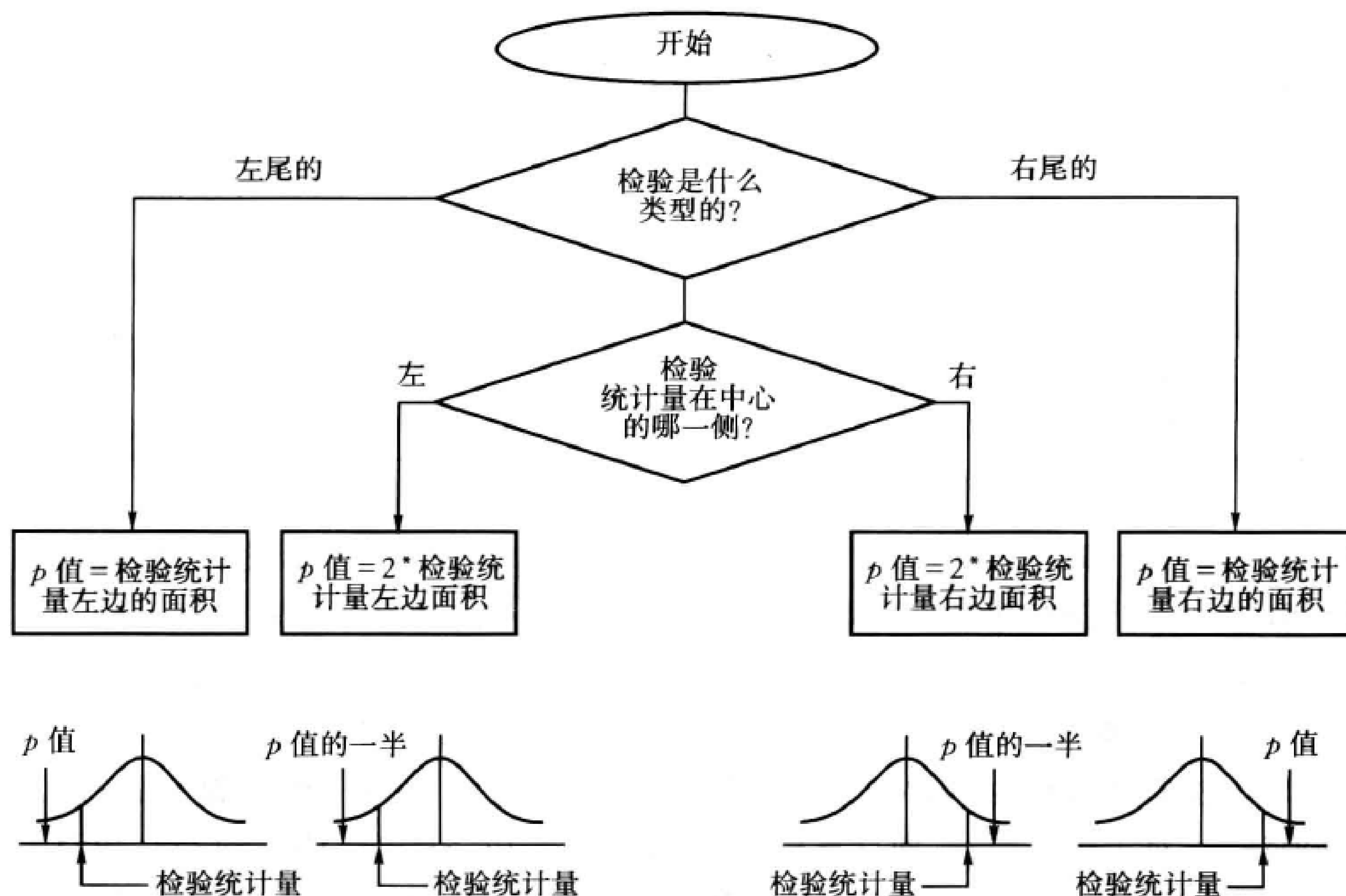


图 6.2.3

进，至于其平均指标是否与 120 有差距的问题，可以再补充一些数据再检验，最好是先采取措施把方差缩小些再决定这个问题。

例 6.2.2 一家食品加工公司的质量管理部门规定，某种包装食品每包净重不得少于 20 千克，经验表明，重量近似服从正态分布，假定得到 50 包食品构成的样本为

19.5，19.0，20.1，21.0，18.9，20.3，21.5，18.8，19.6，19.8，

19.8，19.6，19.6，18.9，17.8，18.0，20.0，20.3，21.0，21.2，

18.5，19.9，20.6，20.1，21.1，22.0，20.8，20.4，20.4，20.3，

19.5，19.5，20.0，21.0，18.9，19.6，19.8，20.0，21.0，20.1，

20.0，18.8，18.9，20.0，21.0，19.6，19.8 19.6，20.0，19.9。

问有无充分证据说明这些包装食品的平均重量减少了？

解 依题意，总体为：包装食品每袋净重量 $X\sim N(\mu,\sigma^2)$，如果把平均重量保持不变或增加作为基本假设的内容，那么只要能拒绝基本假设，就能说明样本数据提供了充分证据说明平均重量减少了。这个理由暗示了应建立如下的假设（单侧检验）

$$H_0: \mu \geqslant \mu_0 = 20, \quad H_1: \mu < \mu_0$$

建立 SPSS 数据文件(表 6.2.5，见 SPSS 数据文件例 6.2.2)，

表 6.2.5　50 包食品重量

	x		x		x
1	19.5	18	20.3	35	18.9
2	19.0	19	21.0	36	19.6
3	20.1	20	21.2	37	19.8
4	21.0	21	18.5	38	20.0
5	18.9	22	19.9	39	21.0
6	20.3	23	20.6	40	20.1
7	21.5	24	20.1	41	20.0
8	18.8	25	21.1	42	18.8
9	19.6	26	22.0	43	18.9
10	19.8	27	20.8	44	20.0
11	19.8	28	20.4	45	21.0
12	19.6	29	20.4	46	19.6
13	19.6	30	20.3	47	19.8
14	18.9	31	19.5	48	19.6
15	17.8	32	19.5	49	20.0
16	18.0	33	20.0	50	19.9
17	20.0	34	21.0	51	

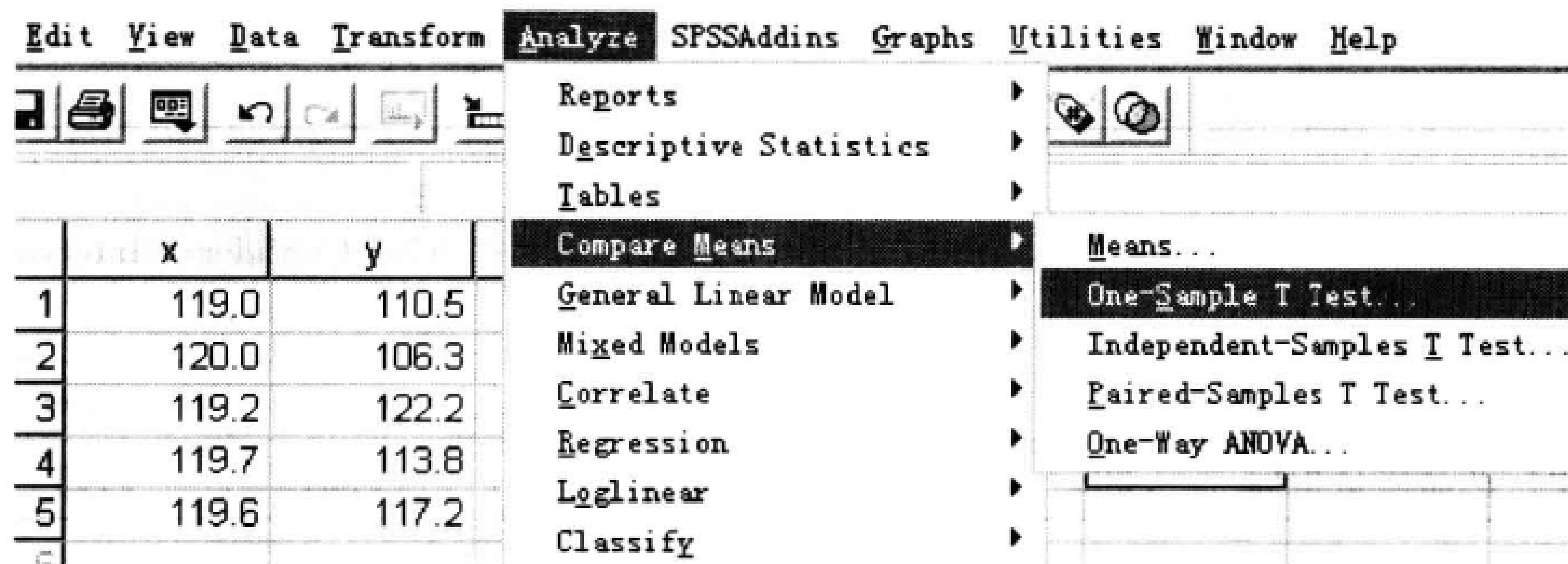


图 6.2.4

在 SPSS 数据编辑窗口，依次调用程序如图 6.2.4 所示。得到对话框并按照需要勾选相应的项(如图 6.2.5)。

再点击 Continue→OK，便得到输出结果如表 6.2.6 和表 6.2.7。

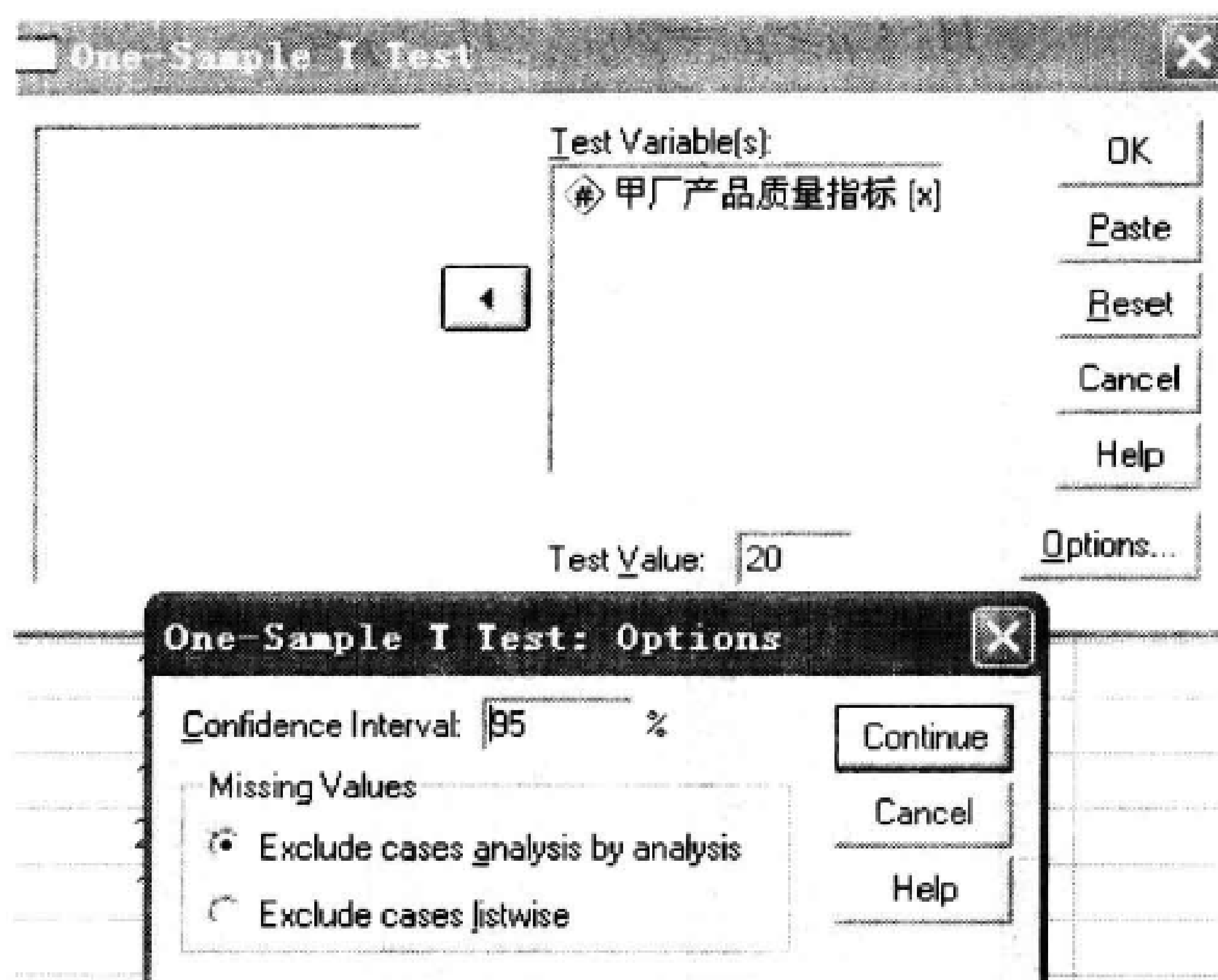


图 6.2.5

表 6.2.6 One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
甲厂产品质量指标	50	19.916	.8674	.1227

表 6.2.7 One-Sample Test

	Test Value = 20					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
甲厂产品质量指标	−.685	49	.497	−.084	−.331	.163

由 $p=\frac{1}{2}\times 0.497=0.2485>\alpha=0.05$，所以，不拒绝基本假设 H_0：$\mu\geqslant\mu_0=20$，即没有充分证据说明这些包装食品的平均重量减少了。

当样本容量超过 30 时，t 检验的临界值(或概值)可以用 z 检验的临界值(或概值)近似代替。

下面看这样一个例子。

研究人员试图找出“客户服务对经理人员很重要”的原因，便调查了制造厂甲的管理领导。提出的原因之一是，客户服务是一种留住顾客的方法。若以 1—5 计

分(1 为低,5 为高),调查回应者对这项原因的评分最高,均值为 4.30。假如制造厂乙的研究人员认为该厂管理者不会给出如此高的评分,为了证明其理论,他们进行了一次假设检验。检验水平设为 0.05,收集数据,得到了以下结果

3　4　5　5　4　5　5　4　4　4　4
4　4　4　4　5　4　4　4　3　4　4
4　3　5　4　4　5　4　4　4　5

利用这些数据和假设检验方法,乙管理者对这项原因的评分是否显著低于在甲得到的均值 4.30?

解　建立假设 $H_0: \mu=4.30$, $H_1: \mu>4.30$

选取检验统计量　$Z=\dfrac{\bar{x}-\mu_0}{\dfrac{s}{\sqrt{n}}}$

检验水平 $\alpha=0.05$　$z_\alpha=1.645$

拒绝域 $W=\{z<-1.645\}$

检验统计量取值　$z=\dfrac{4.156-4.30}{\dfrac{0.574}{\sqrt{32}}}=-1.42$

因为$-1.42>-1.645$,故不能拒绝原假设。即检验产生的数据并不足以表明:客户服务作为一种留住顾客的方式,在乙眼中的重要性要低于它在甲管理者眼中的重要性。对管理者来说,客户服务在甲、乙两厂都是一种留住顾客的重要手段。

使用概值法,检验统计量的观察值 $z=-1.42$,当原假设为真时,取得不大于-1.42 的 z 值的概率为 0.0778,所以在 $\alpha=0.05$ 水平下,不能拒绝原假设。

此问题的样本容量大于 30,使用样本标准差 0.574 代替总体标准差,我们使用 z 检验而不用 T 检验。

2. 方差的检验

与单正态总体均值的检验相比,方差的检验在应用上较少一些,但也有一些应用。比如,一种仪器或一种测定方法的精度(指其内在误差,不是指由于没有调准而产生的偏离)是否达到某种界限,当一种产品的质量问题主要在于波动太大时,可能需要检验方差。

现假定总体 $X\sim N(\mu,\sigma^2)$, $X_1, X_2, \cdots, X_n$为取自该总体的样本, σ^2 未知, μ 可以未知或已知。以下我们分别进行介绍。

(1) μ 未知的情况

建立假设 $H_0: \sigma^2=\sigma_0^2$, $H_1: \sigma^2\neq\sigma_0^2$($\sigma_0^2$ 为已知常数),考虑统计量 $\chi^2=$

$\frac{(n-1)S^2}{\sigma_0^2}$，其中 S^2 为样本方差，它是 σ^2 的无偏估计量。当 $H_0: \sigma^2=\sigma_0^2$ 成立时，$\chi^2=\frac{(n-1)S^2}{\sigma_0^2}\sim\chi^2(n-1)$。于是对给定的显著性水平 α，由

$$P(\chi^2_{1-\frac{\alpha}{2}}(n-1)\leqslant\chi^2=\frac{(n-1)S^2}{\sigma_0^2}\leqslant\chi^2_{\frac{\alpha}{2}}(n-1))=1-\alpha$$

可知 $H_0: \sigma^2=\sigma_0^2$ 的拒绝条件为

$$\chi^2<\chi^2_{1-\frac{\alpha}{2}}(n-1) \text{ 或 } \chi^2>\chi^2_{\frac{\alpha}{2}}(n-1)$$

上述检验法应用了服从 χ^2 分布的检验统计量，姑且称为 χ^2 检验法。

对于单侧假设

$$H'_0: \sigma^2\leqslant\sigma_0^2,\ H'_1: \sigma^2>\sigma_0^2$$

仍选用检验统计量 $\chi^2=\frac{(n-1)S^2}{\sigma_0^2}$，$H'_0: \sigma^2\leqslant\sigma_0^2$的拒绝条件为

$$\chi^2>\chi^2_{\alpha}(n-1)$$

类似地，对于单侧假设

$$H''_0: \sigma^2\geqslant\sigma_0^2,\ H''_1: \sigma^2<\sigma_0^2$$

拒绝 $H''_0: \sigma^2\geqslant\sigma_0^2$ 的条件为

$$\chi^2<\chi^2_{1-\alpha}(n-1)$$

例 6.2.3 在正常的生产条件下，某产品的测试指标总体 $N\sim N(\mu_0, \sigma_0^2)$，其中 $\sigma_0=0.23$。后来改变了生产工艺，出了新产品，假设新产品的测试指标总体仍为 X，且知 $X\sim N(\mu, \sigma^2)$。从新产品中随机地抽取 10 件，测得样本值为 x_1，x_2，…，x_{10}计算得到样本标准差 $S=0.33$。试在检验水平 $\alpha=0.05$ 的情况下检验:① 方差 σ^2 有没有显著变化? ② 方差 σ^2 是否变大?

解 这是单正态总体在总体均值 μ 未知的情况下，方差 σ^2 的统计检验问题。①是双侧检验，② 是单侧检验。

①建立假设 $H_0: \sigma^2=\sigma_0^2=0.23^2$，$H_1: \sigma^2\neq\sigma_0^2$

$H_0: \sigma^2=\sigma_0^2$ 的拒绝条件为

$$\chi^2<\chi^2_{1-\frac{\alpha}{2}}(n-1) \text{ 或 } \chi^2>\chi^2_{\frac{\alpha}{2}}(n-1)$$

因为 $\chi^2=\frac{(n-1)S^2}{\sigma_0^2}=\frac{(10-1)\times 0.33^2}{0.23^2}=18.5274$，$\chi^2_{\frac{\alpha}{2}}(n-1)=\chi^2_{0.025}(9)=19.02$，

$$\chi^2_{1-\frac{\alpha}{2}}(n-1)=\chi^2_{0.975}(9)=2.70$$

而 $2.70<18.5274<19.02$，所以不能拒绝 $H_0: \sigma^2=\sigma_0^2=0.23^2$，即新产品指标的方差与正常情况下产品指标的方差比较没有显著变化。

② 建立假设 $H'_0: \sigma^2\leqslant\sigma_0^2=0.23^2$，$H'_1: \sigma^2>\sigma_0^2$

因为 $\chi^2_{\alpha}(n-1)=\chi^2_{0.05}(9)=\text{IDF.CHISQ}(0.95, 9)=16.92$，此时 χ^2 检验统计

量值仍为 18.5274，又 18.5274>16.92，所以要拒绝 $H'_0: \sigma^2 \leqslant \sigma_0^2 = 0.23^2$，而不拒绝 $H'_1: \sigma^2 > \sigma_0^2$，说明新产品指标的方差比正常情况下产品指标的方差显著地变大。

注　本例中①、②两种情况下的结论好像是矛盾的：①中说没有显著的变化，②中说显著地变大。这是为什么呢？这是因为任何一个统计检验都是在一定的检验水平 α 下进行的，对同一个 α，不同的假设有着不同的拒绝条件（或不拒绝条件）。由于①、②是不同的假设检验问题，拒绝条件不同。因此，同一个 χ^2 值，不满足①的拒绝条件却满足②的拒绝条件，也就没什么奇怪的了。

(2) μ 已知的情形

$H_0: \sigma^2 = \sigma_0^2$，$H_1: \sigma^2 \neq \sigma_0^2$；$H'_0: \sigma^2 \leqslant \sigma_0^2$，$H'_1: \sigma^2 > \sigma_0^2$；$H''_0: \sigma^2 \geqslant \sigma_0^2$，$H''_1: \sigma^2 < \sigma_0^2$。

所有概念上的讨论与前一段没有本质差异，选用检验统计量为

$$\chi^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma_0^2}$$

当 $H_0: \sigma^2 = \sigma_0^2$ 成立时，$\chi^2 = \dfrac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma_0^2} \sim \chi^2(n)$，类似地可分别写出它们的拒绝条件。下面小结如下。

表 6.2.8　单正态总体方差检验表

H_0	H_1	μ 未知时 检验统计量 $\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$ （χ^2 检验）	μ 已知时 检验统计量 $\chi^2 = \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma_0^2}$ （χ^2 检验）
		在显著性水平 α 下拒绝 H_0 的条件	
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\chi^2 < \chi^2_{1-\frac{\alpha}{2}}(n-1)$ 或 $\chi^2 > \chi^2_{\frac{\alpha}{2}}(n-1)$	$\chi^2 < \chi^2_{1-\frac{\alpha}{2}}(n)$ 或 $\chi^2 > \chi^2_{\frac{\alpha}{2}}(n)$
$\sigma^2 \leqslant \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\chi^2 > \chi^2_{\alpha}(n-1)$	$\chi^2 > \chi^2_{\alpha}(n)$
$\sigma^2 \geqslant \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\chi^2 < \chi^2_{1-\alpha}(n-1)$	$\chi^2 < \chi^2_{1-\alpha}(n)$

练习 6.2

1. 设 $X_1, X_2, \cdots, X_n$ 为取自总体 $N(\mu, \sigma^2)$ 的样本，参数 μ, σ^2 均未知，且 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i, Q=\sum_{i=1}^{n}(X_i-\overline{X})^2$，则假设 $H_0: \mu=0$ 的 t－检验，使用的统计量 $T=$ ＿＿＿＿＿＿。

2. 设某厂生产的一种钢索，其断裂强度 X 千克/平方厘米服从正态分布 $N(\mu, 40^2)$。从中选取一个容量为 9 的样本，得 $\overline{X}=780$ 千克/平方厘米。能否据此认为这批钢索的断裂强度为 800 千克/平方厘米($\alpha=0.05$)。

3. 食品厂用自动装罐机装罐头食品，每罐标准重量为 500 克，每隔一定时间需要检验机器的工作情况，现抽 10 罐，测得其重量(单位：克)

495, 510, 505, 498, 503, 492, 502, 512, 497, 506

假设重量 X 服从正态分布 $N(\mu, \sigma^2)$，试问机器工作是否正常($\alpha=0.02$)?

4. 某厂对废水进行处理，要求某种有害物质的浓度不超过 19 毫克/立升，抽样检测得到 10 个数据，其样本均值 $\overline{X}=19.5$ 毫克/立升，样本方差 $S^2=1.25$ (毫克/立升)^2。问在显著性水平 $\alpha=0.05$ 下能认为处理后的废水符合标准吗?

5. 用过去的铸造方法，零件强度服从正态分布，其标准差为 1.6 千克/平方毫米。为了降低成本，改变了铸造方法，测得用新方法铸出零件强度如下：

51.9, 53.0, 52.7, 54.1, 53.2, 52.3, 52.5, 51.1, 54.7

问改变方法后零件强度的方差是否发生了显著变化(取显著性水平 $\alpha=0.05$)?

6. 用包装机包装某种洗衣粉，在正常情况下，每袋重量为 1000 克，标准差不能超过 15 克。假设每袋洗衣粉的重量服从正态分布。某天检验机器工作的情况，从已装好的袋中随机抽取 10 袋，测得其重量(单位：克)为

1020, 1030, 968, 994, 1014, 998, 976, 982, 950, 1048

问这天机器是否工作正常($\alpha=0.05$)?

7. 利用已给出的数据检验下面的假设，并求概值 p

(1) $H_0: \mu=25 \quad vs \quad H_1: \mu\neq 25$

$\bar{x}=28.1$, $n=57$, $s=8.46$, $\alpha=0.01$

(2) $H_0: \mu=7.48 \quad vs \quad H_1: \mu<7.48$

$\bar{x}=6.91$, $n=96$, $s=1.21$, $\alpha=0.01$

(3) $H_0: \mu=1200 \quad vs \quad H_1: \mu>1200$

$\bar{x}=1215$, $n=113$, $s=100$, $\alpha=0.01$

8. 某城市用水工程学会估计，在该城市每户每天的平均用水量为123(单位)，假设有些研究人员认为现在用水量增加了，并想通过检验确定是否确实如此。他们抽取了该城市一些居民作为样本，并细致地记录下某一天中每个样本成员用掉的水，然后使用一种统计软件包进行了分析，结果如下。

One－Sample Statistics

变量	N	Mean	Std. Deviation	Std. Error Mean
用水量	40	132.36	27.68	4.38

One-Sample Test

变量	Test of μ= 123 vs μ>123		
	z	Sig. (2-tailed)	95% Lower Bound
用水量	2.14	0.032	125.16

假定检验水平为0.05，样本容量是多少？样本均值与样本标准差是多少？是单侧检验还是双侧检验？研究结果是什么？根据结果，可以对原假设作出什么样的判断？

6.3　双正态总体的统计检验及SPSS实现

1. 双总体均值之差的检验

设总体 $X\sim N(\mu_1, \sigma_1^2)$，总体 $Y\sim N(\mu_2, \sigma_2^2)$，从 $X\sim N(\mu_1, \sigma_1^2)$ 中抽取样本 $X_1, X_2, \cdots, X_{n_1}$，从 $Y\sim N(\mu_2, \sigma_2^2)$ 中抽取样本 $Y_1, Y_2, \cdots, Y_{n_2}$，且假定两样本间相互独立。

给定常数 μ_0，所要考虑的统计假设有

$$H_0: \mu_1-\mu_2=\mu_0, \quad H_1: \mu_1-\mu_2\neq\mu_0$$
$$H'_0: \mu_1-\mu_2\leqslant\mu_0, \quad H'_1: \mu_1-\mu_2>\mu_0$$
$$H''_0: \mu_1-\mu_2\geqslant\mu_0, \quad H''_1: \mu_1-\mu_2<\mu_0$$

应用上常见的情况是 $\mu_0=0$，对于总体方差 σ_1^2, σ_2^2，我们依下面三种情况分别进行研究。

(1) σ_1^2，σ_2^2，均已知

选取检验统计量 $Z=\dfrac{\overline{X}-\overline{Y}-\mu_0}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}$，当 $H_0: \mu_1-\mu_2=\mu_0$ 成立时，$Z\sim N(0,1)$，于是对给定的显著性水平 α，

$H_0: \mu_1-\mu_2=\mu_0$ 的拒绝条件为：$|Z|>z_{\frac{\alpha}{2}}$

$H'_0: \mu_1-\mu_2\leqslant\mu_0$ 的拒绝条件为：$Z>z_\alpha$

$H''_0: \mu_1-\mu_2\geqslant\mu_0$ 的拒绝条件为：$Z<-z_\alpha$

例 6.3.1 从两个教学班各随机选取 14 名学生进行数学测验，第一教学班与第二教学班测验结果分别由表 6.3.1 中的 A 列与 B 列单元格所示，已知两教学班数学成绩的方差分别为 57 与 53，在显著性水平 0.05 下，可否认为这两个教学班学生的数学测验成绩有差异？

表 6.3.1

	A	B
1	第一班	第二班
2	91	90
3	80	91
4	76	80
5	98	92
6	95	92
7	92	94
8	90	96
9	91	93
10	80	95
11	92	69
12	100	90
13	92	92
14	98	94
15	98	96

解 设第一教学班数学成绩 $X\sim N(\mu_1,57)$，第二教学班数学成绩 $Y\sim N(\mu_2,53)$，$n_1=n_2=n=14$，$\alpha=0.05$。

建立假设 $H_0: \mu_1-\mu_2=0$，$H_1: \mu_1-\mu_2\neq0$

检验统计量 $Z=\dfrac{\overline{X}-\overline{Y}-\mu_0}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}=\dfrac{90.93-90.29-0}{\sqrt{\dfrac{57}{14}+\dfrac{53}{14}}}=0.2293$，$z_{\frac{\alpha}{2}}=z_{0.025}=1.96$，因为 $|0.2293|<1.96$，不满足拒绝 $H_0: \mu_1-\mu_2=0$ 的条件，故不拒绝 $H_0: \mu_1-\mu_2=0$，表

示无充分证据显示两个教学班数学成绩有差异。

(2) σ_1^2，σ_2^2 **均未知，但** $\sigma_1^2=\sigma_2^2=\sigma^2$

选取检验统计量 $T=\dfrac{\overline{X}-\overline{Y}-\mu_0}{S_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$，其中 $S_p^2=\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$，$S_1^2$，$S_2^2$ 分别为两样本的样本方差。当 $H_0: \mu_1-\mu_2=\mu_0$ 成立时，$T\sim t(n_1+n_2-2)$，于是对给定的显著性水平 α，

$$H_0: \mu_1-\mu_2=\mu_0 \text{的拒绝条件为：} |T| > t_{\frac{\alpha}{2}}(n_1+n_2-2)$$

$$H'_0: \mu_1-\mu_2\leqslant\mu_0 \text{的拒绝条件为：} T > t_\alpha(n_1+n_2-2)$$

$$H''_0: \mu_1-\mu_2\geqslant\mu_0 \text{的拒绝条件为：} T > -t_\alpha(n_1+n_2-2)$$

例 6.3.2　某地区高考负责人想知道能不能说某年来自城市中学考生的平均成绩比来自农村中学考生的平均成绩高，已知总体服从正态分布且方差大致相同，由抽样获得资料如表 6.3.2 中 A 列和 B 列。

解　建立假设：$H_0: \mu_1-\mu_2\leqslant 0$，$H_1: \mu_1-\mu_2>0$。

建立 SPSS 数据文件(表 6.3.3，见 SPSS 数据文件例 6.3.2)。第一列 x 为每个考生高考平均成绩，第二列分组变量 g 中“1”代表城市考生，“2”代表农村考生。

表 6.3.2

	A	B
1	城市	农村
2	85	88
3	75	78
4	92	91
5	78	83
6	88	92
7	94	96
8	85	88
9	89	97
10	78	83
11	91	93
12		

表 6.3.3

	x	g
1	85.0	1
2	75.0	1
3	92.0	1
4	78.0	1
5	88.0	1
6	94.0	1
7	85.0	1
8	89.0	1
9	78.0	1
10	91.0	1
11	88.0	2
12	78.0	2
13	91.0	2
14	83.0	2
15	92.0	2
16	96.0	2
17	88.0	2
18	97.0	2
19	83.0	2
20	93.0	2

依次调出双正态总体独立样本 T 检验程序(如图 6.3.1)。

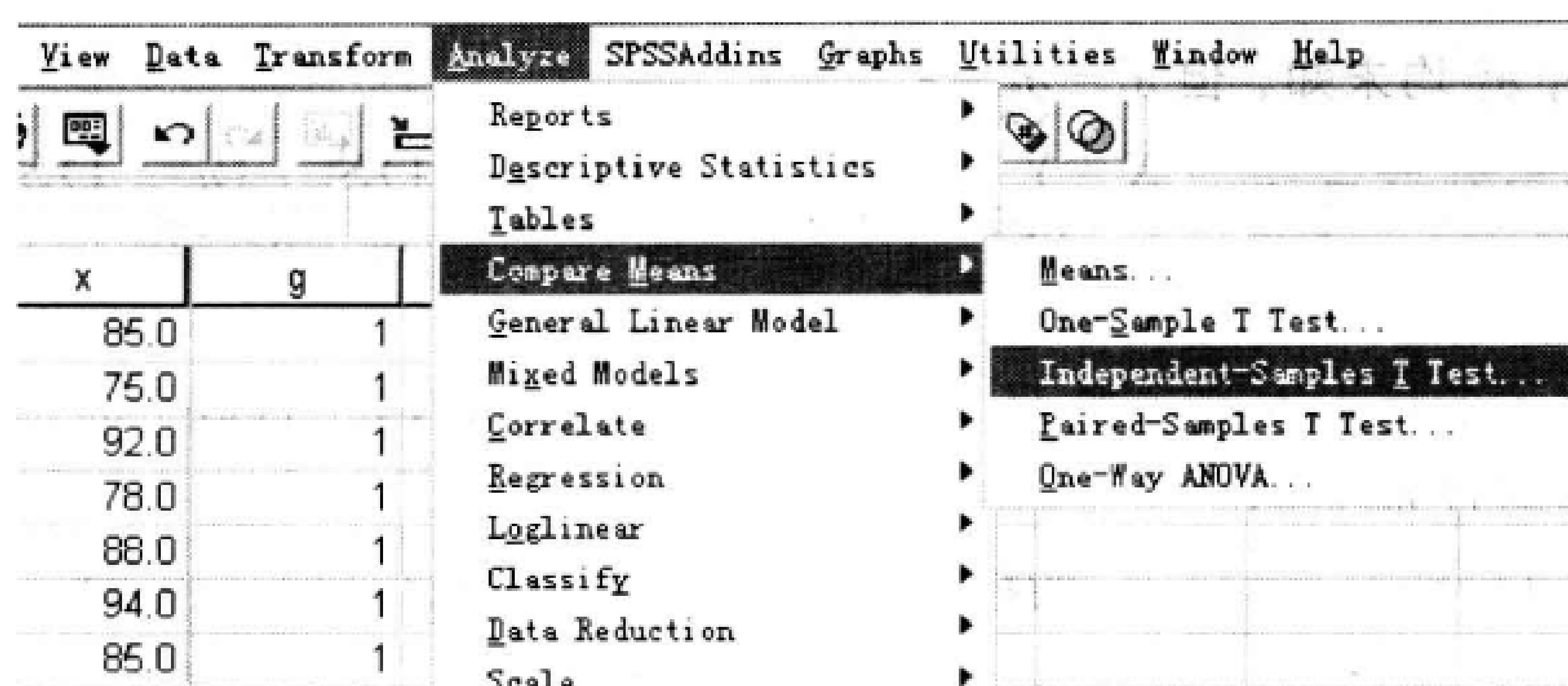


图 6.3.1

打开对话框(如图 6.3.2)。点击 Continue→OK,即得到输出结果(表 6.3.4 和表 6.3.5)。

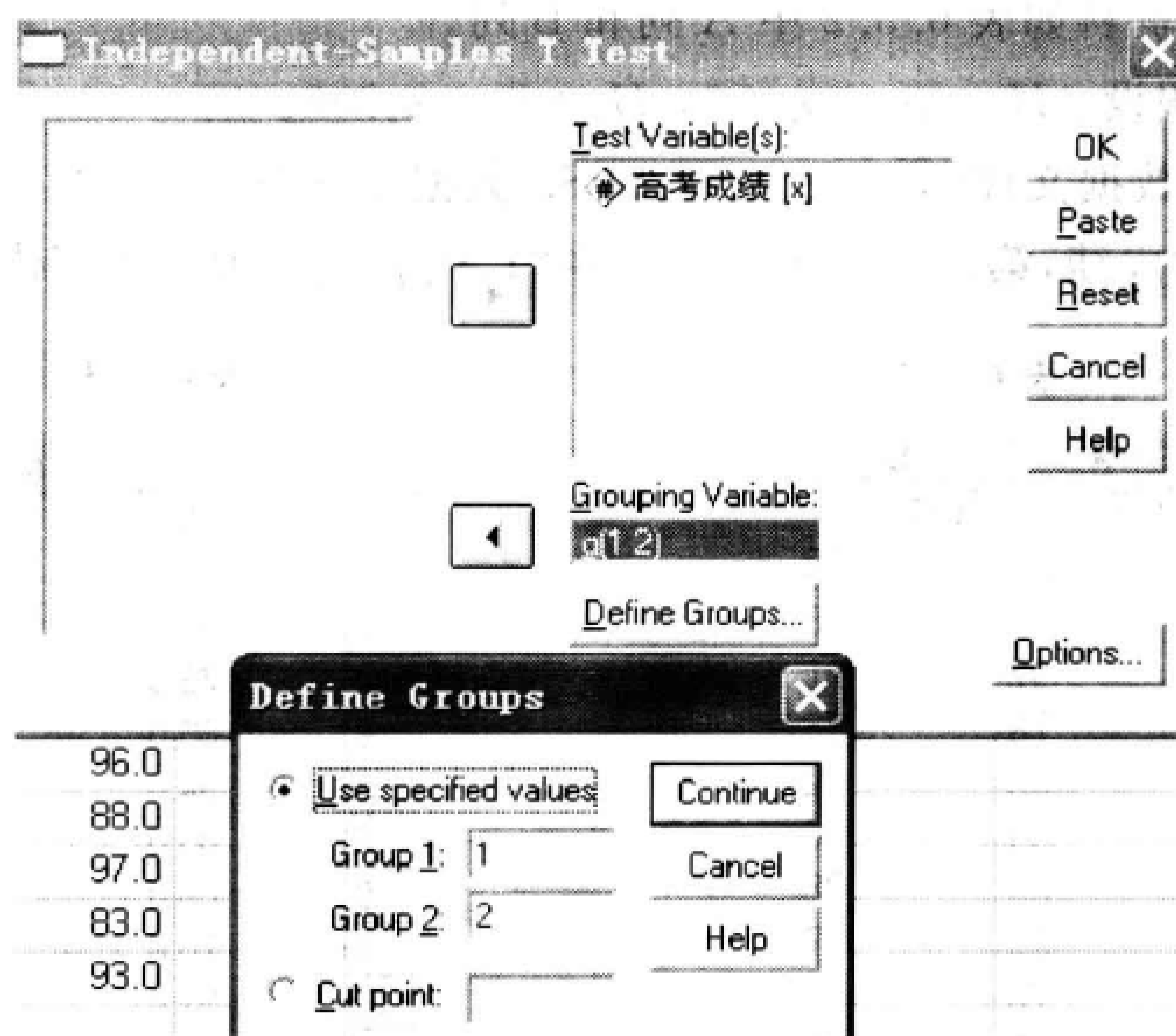


图 6.3.2

表 6.3.4 Group Statistics

	分组	N	Mean	Std. Deviation	Std. Error Mean
高考成绩	城市	10	85.500	6.5532	2.0723
	农村	10	88.900	6.1183	1.9348

表 6.3.5 Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
高考成绩	Equal variances assumed	.071	.793	−1.199	18	.246	−3.400	2.8351
	Equal variances not assumed			−1.199	17.916	.246	−3.400	2.8351

因为 $p=\frac{1}{2}\times 0.246=0.123>0.05=\alpha$，所以不拒绝 $H_0: \mu_1-\mu_2\leqslant 0$，表示无充分证据显示来自城市中学考生的平均成绩比来自农村中学考生的平均成绩高。

(3) σ_1^2，σ_2^2 均未知，但 $\sigma_1^2\neq\sigma_2^2$

选取检验统计量 $T=\dfrac{\overline{X}-\overline{Y}-\mu_0}{\sqrt{\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}}}$，当 $H_0: \mu_1-\mu_2=\mu_0$ 成立时，T 近似服从 $t(\mathrm{d}f)$，其中 $\mathrm{d}f=\dfrac{\left(\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}\right)^2}{\dfrac{S_1^4}{n_1^2(n_1-1)}+\dfrac{S_2^4}{n_2^2(n_2-1)}}$，于是对给定的显著性水平 α，不难写出各基本假设的拒绝条件。

比如，已知某市南区与北区近 12 个月降水量，要求以显著性水平 0.01 检验假设：南区雨量不超过北区雨量。假定观测值均取自正态总体，但其方差不等，就是这样的例子。

2. 成对数据比较检验法

下面我们通过三个实例来解释本段要讨论的主题。

实例 1 设某一种农作物有两个品种 A、B，要比较谁的平均亩产量大，按前一段所讨论的检验两个正态总体均值之差的方法，我们可以准备 n_1+n_2 块形状面积相同的地块，其中 n_1 块种植品种 A，得亩产量 $X_1, X_2, \cdots, X_n$，另 n_2 块种植品种 B，得亩产量 $Y_1, Y_2, \cdots, Y_{n_2}$，然后按上段检验法去处理。这样做有一个前提，就是这 n_1+n_2 个地块的条件必须比较一致。不然的话，假如分配给品种 A

的那 n_1 块地比较肥沃，或其它条件较好，则即使 A 品种并不优于 B，试验结果也可能有利于它。改进的方法是取 n 对地块，每对包含两个形状条件一致的地块，其中一块种植 A，另一块种植 B（哪一块给 A 可随机决定）。这样设计时，哪一个品种也不会占地利之便，不同对的地块条件不必一致，因而较容易办到。

实例 2　为治疗某种疾病，以往用一种药品 A，现新研制出一种药品 B。为比较 A、B 的效果，可以取 n_1+n_2 个患者，n_1 个用 A，n_2 个用 B。但这样做又有与实例 1 相似的问题：患者的情况不一，有的病情已重，一般身体条件差，用药难以见效，有的患者则条件好些。为了避免这种误差，我们可以取 n 对患者，每对的两名患者在条件上尽可能一致，其中一名用 A，另一名用 B，不同对患者条件不必一致。这样做避免了上述误差，在设计上也不难实现。因为这里只要求每对的两名患者条件一致，而不要求全体参试的患者都一致。

实例 3　为比较每天在固定时间上八小时班和让工作者自己选择八小时工作时间这两种方式哪一种更有效(能完成更多的工作量)。可以挑选 n_1+n_2 个工作者作试验，其中 n_1 个采用第一种方式，余下的 n_2 个采用第二种方式。但这样做又有一个由于这 n_1+n_2 个工作者工作能力高低不同而带来的误差问题。为克服这一点，我们采取成对比较法，选择 n 对工作者，每对包含条件尽可能接近的两名工作者，这种试验设计(即成对设计)就避免了上述误差。

总结以上诸例的思想，我们就可以提出以下的一般模型：设有两个需要进行比较的处理，“处理”一词的含义在此很广泛，如在实例 1 中，每个品种是一个处理；在实例 2 中，每种药物是一个处理；在实例 3 中，每种工作方式是一个处理，等等。选择 n 对“试验单元”，每对中的两个试验单元条件尽可能一致，而不同对之间则不要求一致，在每一对内，随机地决定把其中的一个试验单元给处理 1，另一个给处理 2，经过试验，观测各处理在每个试验单元上的试验结果，如下表 6.3.6。

表 6.3.6

对	处理 1	处理 2	差 $Z_i=X_i-Y_i$
1	X_1	Y_1	$Z_1=X_1-Y_1$
2	X_2	Y_2	$Z_2=X_2-Y_2$
⋮	⋮	⋮	⋮
n	X_n	Y_n	$Z_n=X_n-Y_n$

这里的 $Z_i=X_i-Y_i$，$i=1, 2, \cdots, n$ 就是在第 i 对试验单元中，所观测到的处理 1 优于处理 2 的量(为确定计，我们假定观测值愈大愈优)。这个量不是由于试

验条件上的差别而来，因为每对内两个试验单元条件已尽量一致了。我们假定 $Z_i=X_i-Y_i$ 服从正态分布 $N(\mu, \sigma^2)$，而 μ 就表示处理 1 平均优于处理 2 的量。这样一来，两种处理的比较就归结为对 μ 的检验问题，例如：

① 两种处理效果一样，$\mu=0$；

② 处理 2 不优于处理 1，$\mu\geqslant 0$；

③ 处理 2 不劣于处理 1，$\mu\leqslant 0$；

④ 处理 1 平均优于处理 2 的量为 μ_0，$\mu=\mu_0$；

⑤ 处理 1 平均优于处理 2 的量不超过 μ_0，$\mu\leqslant\mu_0$；

⑥ 处理 1 平均优于处理 2 的量不小于 μ_0，$\mu\geqslant\mu_0$

等等。因此，问题回到我们已讨论过的单正态总体的均值的检验。

$Z_1, Z_2, \cdots, Z_n$ 可视为取自正态总体 $N(\mu, \sigma^2)$ 的样本，所以，我们可选取检验统计量

$$T=\frac{\bar{Z}-\mu_0}{\frac{S_d}{\sqrt{n}}}$$

其中 $\bar{Z}=\bar{X}-\bar{Y}$，$S_d^2=\frac{1}{n-1}\sum_{i=1}^{n}(Z_i-\bar{Z})^2$，当 $H_0: \mu=\mu_0$ 成立时，$T\sim t(n-1)$，对给定的显著性水平 α，

$H_0: \mu=\mu_0$ 的拒绝条件为：$|T|>t_{\frac{\alpha}{2}}(n-1)$

$H'_0: \mu\leqslant\mu_0$ 的拒绝条件为：$T>t_\alpha(n-1)$

$H''_0: \mu\geqslant\mu_0$ 的拒绝条件为：$T<-t_\alpha(n-1)$

例 6.3.3　为了解两种教学法对 9 名学生试验的结果，经试验后，测得成绩如表 6.3.7 的 A 列和 B 列。假定总体为正态，以 0.05 为显著性水平，检验此两种教学法效果是否不同？

表 6.3.7

	A	B
1	方法A	方法B
2	327.6	327.6
3	327.7	327.7
4	327.7	327.6
5	327.9	327.8
6	327.4	327.4
7	327.7	327.6
8	327.8	327.8
9	327.8	327.7
10	327.4	327.3

解　检验原假设 $H_0: \mu=\mu_0=0$，使用 SPSS 解成对观测值 t 检验方法如下：

建立 SPSS 数据文件(表 6.3.8,见 SPSS 数据文件例 6.3.3)。

表 6.3.8

	x	y
1	327.6	327.6
2	327.7	327.7
3	327.7	327.6
4	327.9	327.8
5	327.4	327.4
6	327.7	327.6
7	327.8	327.8
8	327.8	327.7
9	327.4	327.3

调用分析程序,依次点击(如图 6.3.3)。

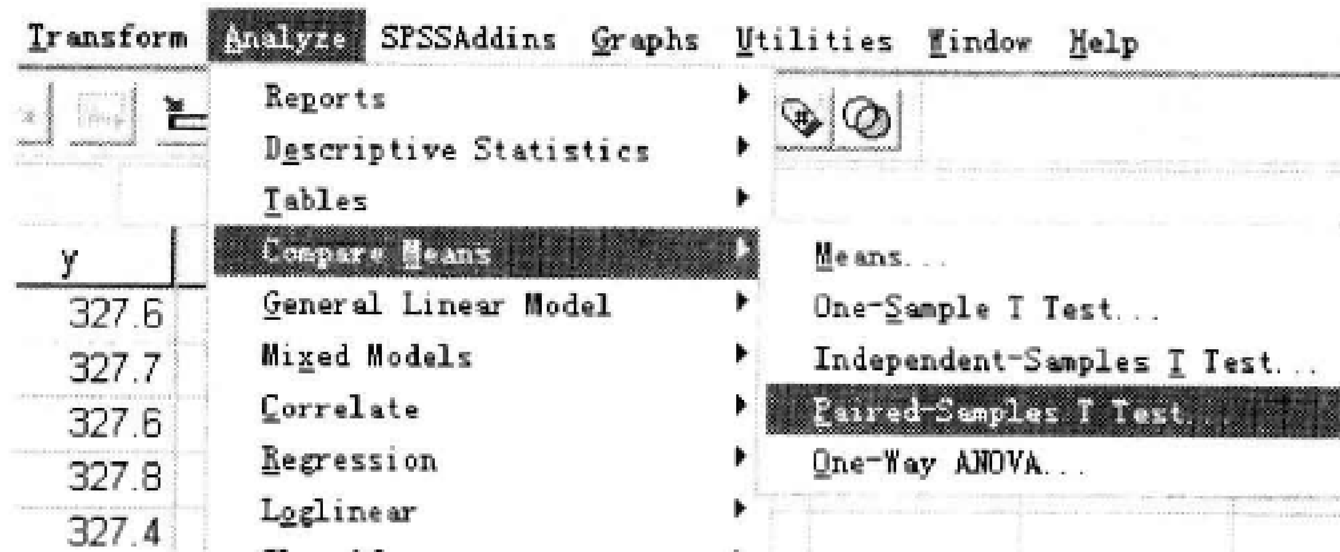


图 6.3.3

弹出对话框并填写相应选项(图 6.3.4)。

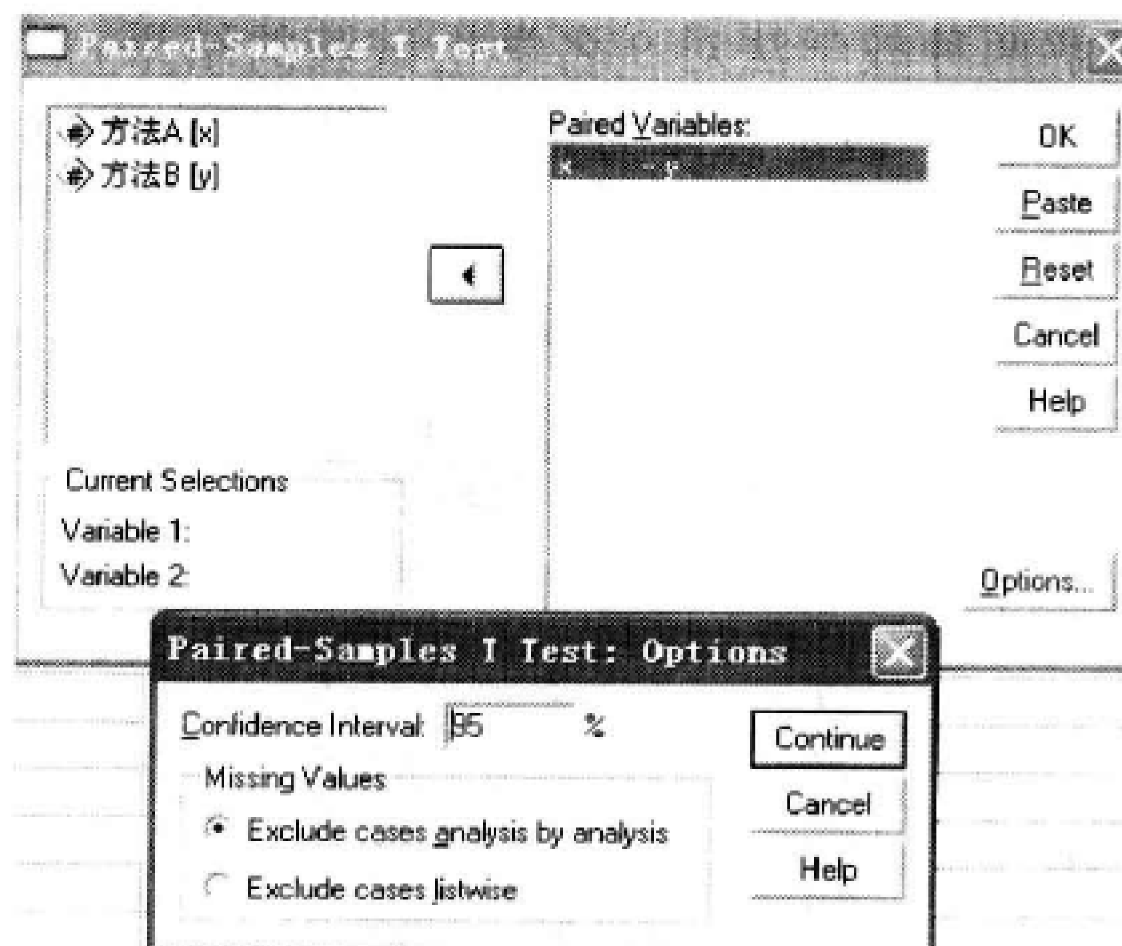


图 6.3.4

点击 Continue→OK，即得到输出结果(表 6.3.9—表 6.3.11)。

表 6.3.9　**Paired Samples Statistics**

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	方法 A	327.667	9	.1732	.0577
	方法 B	327.611	9	.1691	.0564

表 6.3.10　**Paired Samples Correlations**

		N	Correlation	Sig.
Pair 1	方法 A & 方法 B	9	.953	.000

表 6.3.11　**Paired Samples Test**

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1 方法 A 方法 B	.056	.0527	.0176	.015	.096	3.162	8	.013

因为 $p=0.013<0.05=\alpha$，故拒绝原假设，而得出此两种教学法效果不同的结论。或 t 值为 3.16227766，自由度为 8，分位数为 2.306005626，因 $3.16227766>2.306005626$，故拒绝原假设，而得出此两种教学法效果不同的结论。

3. 双总体方差之比的检验

设总体 $X\sim N(\mu_1,\sigma_1^2)$，总体 $Y\sim N(\mu_2,\sigma_2^2)$，从 $X\sim N(\mu_1,\sigma_1^2)$ 中抽取样本 $X_1,X_2,\cdots,X_{n_1}$，从 $Y\sim N(\mu_2,\sigma_2^2)$ 中抽取样本 $Y_1,Y_2,\cdots,Y_{n_2}$，且假定两样本间相互独立。

在第 4 章中，我们已经知道

$$F=\frac{\frac{1}{\sigma_1^2}S_1^2}{\frac{1}{\sigma_2^2}S_2^2}\sim F(n_1-1,n_2-1),$$

$$\widetilde{F}=\frac{\frac{1}{\sigma_1^2}\sum_{i=1}^{n_1}(X_i-\mu_1)^2}{\frac{1}{\sigma_2^2}\sum_{i=1}^{n_2}(Y_i-\mu_2)^2}\sim F(n_1,n_2)$$

下面我们讨论双正态总体方差齐性的检验。

提出统计假设

$$H_0:\frac{\sigma_1^2}{\sigma_2^2}=1,\ H_1:\frac{\sigma_1^2}{\sigma_2^2}\neq 1$$

$$H'_0:\frac{\sigma_1^2}{\sigma_2^2}\leqslant 1,\ H'_1:\frac{\sigma_1^2}{\sigma_2^2}>1$$

$$H''_0:\frac{\sigma_1^2}{\sigma_2^2}\geqslant 1,\ H''_1:\frac{\sigma_1^2}{\sigma_2^2}<1$$

(1) μ_1，μ_2 均未知

当 $H_0:\frac{\sigma_1^2}{\sigma_2^2}=1$ 成立时，检验统计量是 $F=\frac{S_1^2}{S_2^2}\sim F(n_1-1,n_2-1)$，对给定的显著性水平 α，由于 $P(F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)\leqslant F\leqslant F_{\frac{\alpha}{2}}(n_1-1,n_2-1))=1-\alpha$，所以，$H_0:\frac{\sigma_1^2}{\sigma_2^2}=1$ 的拒绝条件为

$$F<F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)\ 或\ F>F_{\frac{\alpha}{2}}(n_1-1,n_2-1)$$

注　一般选较大的样本方差做检验统计量的分子，因而也仅需要查分位数 $F_{\frac{\alpha}{2}}(n_1-1,n_2-1)$ 即可得出结果。

类似可得

$$H'_0:\frac{\sigma_1^2}{\sigma_2^2}\leqslant 1\ 的拒绝条件为：F>F_{\alpha}(n_1-1,n_2-1)$$

$$H''_0:\frac{\sigma_1^2}{\sigma_2^2}\geqslant 1\ 的拒绝条件为：F<F_{1-\alpha}(n_1-1,n_2-1)$$

注意到公式 $F_{1-\alpha}(n_1-1,n_2-1)=\frac{1}{F_{\alpha}(n_1-1,n_2-1)}$有时是有用的。

(2) μ_1，μ_2 均已知

当 $H_0:\frac{\sigma_1^2}{\sigma_2^2}=1$ 成立时，检验统计量 $\widetilde{F}=\frac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2}{\sum_{i=1}^{n_2}(Y_i-\mu_2)^2}\sim F(n_1,n_2)$，完全类似于上段，对给定的显著性水平 α，易写出

$$H_0:\frac{\sigma_1^2}{\sigma_2^2}=1\ 的拒绝条件为：\widetilde{F}<F_{1-\frac{\alpha}{2}}(n_1,n_2)\ 或\ \widetilde{F}>F_{\frac{\alpha}{2}}(n_1,n_2)$$

类似可得

$$H'_0: \frac{\sigma_1^2}{\sigma_2^2} \leqslant 1 \text{ 的拒绝条件为：} \widetilde{F} > F_\alpha(n_1, n_2)$$

$$H''_0: \frac{\sigma_1^2}{\sigma_2^2} \geqslant 1 \text{ 的拒绝条件为：} \widetilde{F} < F_{1-\alpha}(n_1, n_2)$$

注意到公式 $F_{1-\alpha}(n_1, n_2) = \dfrac{1}{F_\alpha(n_2, n_1)}$有时是有用的。

练习 6.3

1. 某化工厂为了提高某种化学药品的得率，提出了两种工艺方案。为了研究哪一种方案好，分别用两种工艺各进行了 10 次试验，数据如下

方案甲得率(%)：68.1，62.4，64.3，64.7，68.4，66.0，65.5，66.7，67.3，66.2

方案乙得率(%)：69.1，71.0，69.1，70.0，69.1，69.1，67.3，70.2，72.1，67.3

假设得率服从正态分布，问方案乙是否能比方案甲显著提高得率(取 $\alpha=0.01$)?

2. 用两种方法研究冰的潜热，样本都取自 −0.72 ℃的冰。用方法 A 做，取样本容量 $n_1=13$，用方法 B 做，取样本容量 $n_2=8$，测得每克冰从 −0.72 ℃变成 0 ℃的水，其中热量的变化数据为

方法 A：79.98，80.04，80.02，80.04，80.03，80.04，80.03，79.97，80.05，80.03，80.02，80.00，80.02；

方法 B：80.02，79.94，79.97，79.98，79.97，80.03，79.95，79.97。

假设两种方法测得数据总体都服从正态分布，试问：

(1) 两种方法测量总体的方差是否相等($\alpha=0.05$)?

(2) 两种方法测量总体的均值是否相等($\alpha=0.05$)?

3. 9 名运动员在初进运动队时和接受一周训练后各进行一次体能测试，测试评分为

运动员	1	2	3	4	5	6	7	8	9
入队时	76	71	57	49	70	69	26	65	59
训练后	81	85	52	52	70	63	33	83	62

假设分数服从正态分布，试在显著性水平 $\alpha=0.05$ 下，判断运动员体能训练效果是否显著?

6.4 两个需要说明的问题

1. 统计检验与区间估计的关系

统计检验和区间估计是两种最重要的统计推断形式，这两者初看好像完全不同，其实两者之间有一定的联系。然而，其结果的解释上也有差别。下面我们只就一个例子来加以说明。

(1) 利用统计检验可建立区间估计，反之亦然

以正态分布期望值的检验和区间估计问题为例。设 X_1，X_2，…，X_n 为取自正态总体 $N(\mu, \sigma^2)$的样本，方差 σ^2 未知，要检验 H_0 ：$\mu=\mu_0$，H_1 ：$\mu\neq\mu_0$。我们知道，对给定的检验水平 α，进行 t 检验，H_0 ：$\mu=\mu_0$ 的拒绝条件为 $\left|\dfrac{\overline{X}-\mu_0}{\frac{S}{\sqrt{n}}}\right|>t_{\frac{\alpha}{2}}(n-1)$，不拒绝条件为 $\left|\dfrac{\overline{X}-\mu_0}{\frac{S}{\sqrt{n}}}\right|\leqslant t_{\frac{\alpha}{2}}(n-1)$，将这个不拒绝条件改写成 $\overline{X}-\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)\leqslant\mu_0\leqslant\overline{X}+\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)$，再把 μ_0 改成 μ，便可得到 μ 的置信度为 $1-\alpha$ 的置信区间。

反之，若我们先确定了 μ 的区间估计 $\overline{X}-\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)\leqslant\mu_0\leqslant\overline{X}+\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)$，则将 μ 改成 μ_0，这样就得到了原假设 H_0 ：$\mu=\mu_0$ 的不拒绝条件，也就得到了 H_0 ：$\mu=\mu_0$的拒绝条件，检验水平为 α。

这个对应关系在其它问题中也存在。作为练习，建议考虑在方差相等(但未知)的条件下，两正态总体均值之差的统计检验和区间估计的问题。

(2) 统计检验和区间估计的结果，在解释上可以有差别

仍就上例来说明。在有了样本之后，我们同时检验假设 H_0 ：$\mu=\mu_0=0$(水平 α)及作 μ 的区间估计(置信度为 $1-\alpha$)。对不同的样本值，以下几种情况都可能出现：

① 不拒绝 H_0 ：$\mu=\mu_0=0$，区间估计为$(-0.001, 0.002)$；② 不拒绝 H_0 ：$\mu=\mu_0=0$，区间估计为$(-1000, 1500)$；③ 拒绝 H_0 ：$\mu=\mu_0=0$，区间估计为$(1000, 2000)$；④ 拒绝 H_0 ：$\mu=\mu_0=0$，区间估计为$(0.001, 0.002)$。

情况①，按统计检验，应不拒绝 H_0 ：$\mu=\mu_0=0$，按区间估计，μ 能取的最大值和最小值都很接近 0，这两者的解释一致。

情况②，按统计检验，应不拒绝 $H_0: \mu=\mu_0=0$，按区间估计，这区间包含 0，即 0 是 μ 的一个可能值，在这一点上与统计检验的结论一致。但细看这一区间，最大可以到 1500，最小可到 -1000，这中间哪一个值都可能。因此，从区间估计的角度看，实在没有多大把握认为 μ 能在 0 附近，这就与统计检验给的结论不大协调了。

情况③，按统计检验，应拒绝 $H_0: \mu=0$，按区间估计，这区间不包含 0，即 0 不看作 μ 的可能值，而且，区间的最小值也有 1000，与 0 相距甚远。故认为 $\mu\neq0$ 很有理由(区间估计的结论加强了统计检验的结论)。

情况④，按统计检验，应拒绝 $H_0: \mu=\mu_0=0$，按区间估计，这区间不包含 0，从这方面看两者一致。可是细看这区间，就发现它整个在 0 附近。因此，实质上可以认为 μ 就是 0，这样区间估计的结论(在实质上)就与统计检验的结论不同。

由此例也看到，统计上的结论一定要注意其实质含义，如只停留在表面上，就有可能被引入歧途。

2. 检验的 p 值

统计检验的可能结论只有两个：不拒绝或拒绝。作出这一或那一结论的根据有多大，则往往不能清楚地显示出来。比如，某市的工业主管部门想判定 A，B 两厂的产品质量谁优谁劣，他委托一个统计学家进行一个统计检验 H_0："A 不优于 B"，当他被告知这假设应不被拒绝时，他只知道作出的结论是 H_0："A 不优于 B"，但作出这一结论的根据有多大？他不可能有一个数量的概念。这是统计检验这种统计推断形式的一个缺点。在上一段，我们曾将统计检验和区间估计做了一个对比，并指出：一般说来，用统计检验作出的结论，不如区间估计那么精细。这一点的根由就在于统计检验这种形式固有的粗糙性。但是，对这些情况可以作补救，方法是引进下面要介绍的"p 值"。

设想一个简单的检验问题：在正态总体 $N(\mu, 1)$ 中抽取样本 $X_1, X_2, \cdots, X_{16}$，要检验假设 $H_0: \mu=0$，$H_1: \mu\neq0$，取检验水平 0.05，拒绝 $H_0: \mu=0$ 的条件为 $|\overline{X}|>0.49$。假设对一组具体的 $X_1, X_2, \cdots, X_{16}$ 有 $\overline{X}=0.48$，这时我们应不拒绝 $H_0: \mu=0$。又假设另一组具体的 $X_1, X_2, \cdots, X_{16}$ 有 $\overline{X}=0.12$，则当然也不拒绝 $H_0: \mu=0$。对这两组样本而言，结论一致(都是不拒绝 $H_0: \mu=0$)。然而，我们会觉得，在后一种场合，作出 $\mu=0$ 的结论根据大一些；而在前一场合，此根据则小一些。为了反映这一点，我们引进 p 值。其定义如下。

设对某一组具体样本，计算出 $\overline{X}=b$，则这组样本的 p 值定义为

$$p=P(|Z|>4|b|),\ Z\sim N(0, 1)$$

解释如下：我们已经实际观测到 $\overline{X}$ 的一个值 b，它与假设值 $\mu=0$ 有偏差 $|b|$，我们问达到这么大或更大的偏离的机会有多大？这就是上式定义的 p。若 p 很大，说明在原假设 $\mu=0$ 之下，得到这么大偏差很平常，不值得奇怪，因而认为 $\mu=0$ 的根据较充分。反之，若 p 很小，则在 $\mu=0$ 之下得到这么大一个偏差很难得，这很有可能意味着 $\mu\neq0$。因此，当 p 很小时，认为 $\mu=0$ 的根据就很不够。总之，p 愈大(小)，认为 $\mu=0$ 的根据就愈足(不足)，当 p 值落到给定的水平 α 之下时，就要拒绝 $\mu=0$ 了。若 $p\geqslant\alpha$，但离 α 很近，则我们虽不能拒绝 $\mu=0$，但对它抱着很怀疑的态度。

比如，用前面的举例来说，两组样本的 p 值分别为

$$p=P(|Z|>1.92)=0.0548,\quad p=P(|Z|>0.48)=0.6312$$

前一种情况，p 值离检验水平 $\alpha=0.05$ 很近，虽不能拒绝 $\mu=0$，但是值得怀疑。后一种情况，出现像 0.12 或更大偏差的可能性在 $\mu=0$ 下为 0.6312，这一可能性很大，不足为奇。

以上分析很自然地推广到一般情况：设有一个原假设 H_0，其拒绝条件为 $|t|>C$，t 为检验统计量。若对一组具体样本计算出统计量 t 之值为 t_0，则这组样本的 p 值是 $p=P(|t|>|t_0||H_0)$，如果拒绝条件为 $t>C$，则 p 值为 $p=P(t>t_0|H_0)$，如果拒绝条件为 $t<C$，则 p 值为 $p=P(t<t_0|H_0)$。

例 6.4.1 从电话公司每月长途电话的账单中，随机抽取 37 张，计算平均费用为 33.15 元，标准差为 21.21 元。假定费用服从正态分布 $N(\mu,\sigma^2)$，σ^2 未知，要检验假设 $H_0:\mu=30$，$H_1:\mu\neq30$，试计算 p 值。

解 因为检验统计量为 $t=\dfrac{\overline{X}-\mu_0}{\frac{S}{\sqrt{n}}}$，在 $H_0:\mu=30$ 下 $t=\dfrac{\overline{X}-\mu_0}{\frac{S}{\sqrt{n}}}$ 服从分布 $t(n-1)$，依样本计算检验统计量 $t=\dfrac{\overline{X}-\mu_0}{\frac{S}{\sqrt{n}}}$ 的值为 $t_0=\dfrac{33.15-30}{\frac{21.21}{\sqrt{37}}}=0.90$，所以

$$\begin{aligned}
p&=P(|t|>|t_0|\ |H_0)=P(|t|>0.90|H_0)\\
&=1-P(-0.90\leqslant t\leqslant0.90|H_0)\\
&=1-[\text{CDF.T}(0.90,37-1)-\text{CDF.T}(-0.90,37-1)]\\
&=1-[0.81-0.19]=0.38
\end{aligned}$$

说明样本支持原假设，故不拒绝原假设。

第7章　方差与协方差分析

在现实的生产和经营管理中，经常要分析各种因素对研究对象某些特征值的影响。比如，在工业生产中往往要考察几种不同原料对产品质量有否明显影响；在农业生产中往往要考察不同品种、肥料种类、施肥量等对某种作物亩产量的影响；在化学实验中往往要分析反应温度、反应时间、原料成分、原料用量等对某种物质得率的影响；在市场营销中往往要分析不同地区、不同时期对某种产品销量的影响，等等。在影响特征值的众多因素中，常常需要分析哪几种因素对生产和销量起显著作用，并需知道起显著作用的因素在何时起最好的影响作用。为此，就必须让这些因素改变各种不同状态进行试验和观察，并对试验结果——数据进行科学的分析。方差分析就是采用数理统计方法对数据进行分析，以鉴别各种因素对研究对象的某些特征值影响大小的一种有效方法。

方差分析是英国大统计学家费歇尔(Fisher)在20世纪20年代创立的，那时他在英国一个农业试验站工作，需要进行许多田间试验，为分析这种试验的结果，他发明了方差分析法，之后此方法被用于其它许多领域，都取得了很大的成功。

7.1　方差分析的基本思想

为方便起见，今后我们把研究对象的特征值，即所考察的试验(其涵义包括调查，收集等)结果(如产品质量、数量、销量、成本等)称为试验指标，简称指标，常用 x 表示。由于试验误差的存在，故“指标”为随机变量。在试验中对所关心的“指标”有影响的、要加以考察而改变状态的原因称为因素，如工业生产中影响产品质量的因素有原材料、工艺条件、工人技术水平等，常用 A,B,C 等大写英文字母表示。因素在试验中所取的各种不同状态称为因素的水平，因素 A 的 r 个水平常用 $A_1,A_2,\cdots,A_r$ 表示，其中 r 称为因素 A 的水平数。若只考察一个因素对指标的影响，这种试验称为单因素试验，相应的方差分析称为单因素方差分析；若一个试验中同时考察两个因素，则相应的试验称为双因素试验，这时所作的方差分析称为双因素方差分析，在多因素试验中要考察的因素多于两个，相应的方差分析称为多因

素方差分析。

实例 1 某食品集团的产品销售覆盖全国，主要分布于 25 个省份，是一个颇受消费者青睐的品牌。集团营业部根据其销售情况，将这 25 个省份划分为东北、华北、东南、西北和中部 5 个销售区域，每个区域由一名销售经理负责。年末将近，各部门经理都在准备年度报告。营业部总经理准备分析一下过去一年里各区域的销售业绩。他从营业部专员手中得到了各省份销售情况，发现东北地区今年对集团的收入贡献不大，销售量(万箱)没有其它地区多，其原因可能在于该地区人口规模比其它地区小，消费习惯存在差异等。当然，这种差异也可能是由于偶然因素。但如果各区域间具有显著差异，则应当引起销售部门的注意，从而进一步研究不同区域的不同特征，进一步进行市场细分，采取适当的营销策略。

在形成报告之前，总经理决定对各地区的销售量进行分析，首先检验各区域间的差异是否由于偶然原因所致，确认各区域销售量之间是否存在明显的差异。

实例 2 某公司为了研究三种内容的广告宣传对某种无季节性的大型机械销售量的影响进行了调查统计。经广告广泛宣传后，按寄回广告上的订购数计算，一年四个季度的销售量(单位：台)为

广告类型	第一季度	第二季度	第三季度	第四季度
A_1	163	176	170	185
A_2	184	198	179	190
A_3	206	191	218	224

表中 A_1 是强调运输方便性的广告，A_2 是强调节省燃料经济性的广告，A_3 是强调噪音低的优良性的广告。试判断：新闻广告的类型对该种机械的销售量是否有显著影响？若影响显著，哪一种广告内容为好？

此例中，新闻广告是所要检验的因素，三种不同的内容可看作是三个水平，因而这是一个单因素三水平的试验。若这三种广告内容的宣传对机械销售量的影响没有显著差异，那么从中采取一种既经济又方便的广告即可；若有显著差异，则希望从中选取一种较优的方案，以便对提高机械的销售量更为有利。

从销售量数据来看，好像不同广告对销售量有一定影响；而在同一种广告下，四个不同季节的销售量也不完全一样，由于该机械无季节性，所以产生这种差异的原因认为是试验过程中各种偶然性(随机性)因素的干扰所致，这一类误差称为试验误差，试验误差的存在使不同广告下销售量的差异必须进行仔细鉴别。究竟这差异是单纯由误差引起的，还是由于广告类型不同而引起的。如果单纯是由误差

引起的，那么我们认为广告的不同类型对销售量没有显著影响，则可简称因素(新闻广告)不显著；如果不同广告下销售量的不同，除了误差影响外，主要是由于广告类型(水平)不同所造成的，那么我们就认为因素的不同水平对销售量有显著影响，简称因素显著。方差分析就是通过对试验结果的分析去判断因素本身及各因素间交互作用对指标是否影响显著的一种统计方法。

7.2　单因素方差分析

单因素方差分析是要判断因素各水平对指标是否有显著影响，归结为判断不同总体是否有相同分布的问题。由于实际中常遇到具有正态分布的总体，同时，在进行方差分析时，除了所关心的因素外，其它条件总是尽可能使其保持一致，这样就可以认为每个总体的方差相同。因而，判断几个总体是否具有相同分布的问题就简化为检验几个具有等方差的正态总体均值是否相等的问题。

1. 数学模型

我们考虑的因素记为 A，假定它有 r 个水平，并对水平 A_i 作了 n_i 次观察，第 i 水平的第 j 次观察为 x_{ij}，这样可得观察资料(见表 7.2.1)

表 7.2.1　观测数据

	1	2	…	n_i
A_1	x_{11}	x_{12}	…	x_{1n_1}
A_2	x_{21}	x_{22}	…	x_{2n_2}
…	…	…	…	…
A_r	x_{r1}	x_{r2}	…	x_{rn_r}

设 $x_{ij}(j=1, 2, 3, \cdots, n_i)$是来自总体 $N(\mu_i, \sigma^2)$的简单随机样本，$i=1, 2, 3, \cdots, r$。目的是检验 $H_0: \mu_1=\mu_2=\cdots=\mu_r$。因为 μ_i 是第 i 水平下试验数据的理论均值，所以 $x_{ij}=\mu_i+\varepsilon_{ij}$，$\varepsilon_{ij}\sim N(0, \sigma^2)$，$i=1, 2, 3, \cdots, r$。记 $\mu=\frac{1}{n}\sum_{i=1}^{r}n_i\mu_i$，$\alpha_i=\mu_i-\mu$，这里 α_i 表示第 i 水平对指标的效应值，它反映了因素第 i 水平对指标的纯作用大小，是除了因素对指标的平均影响后，因素的第 i 水平对指标的特殊影响。于是 $x_{ij}=\mu+\alpha_i+\varepsilon_{ij}$，其中 $\sum_{i=1}^{r}n_i\alpha_i=0$，$H_0: \mu_1=\mu_2=\cdots=\mu_r=\mu$ 等价于

$H'_0: \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$。从效应的角度讲，方差分析要解决的问题是：检验因素各水平对试验指标的影响是否显著；若存在显著的差异，表明因素各水平的效应不完全相同，可以从中选出效应最大的水平即最优水平作为实施方案；若无显著差异，表明因素各水平对试验指标的影响一样，差异是随机引起的，这时可以从中选择支出最少或费用最低的水平作为实施方案。

记总观察次数 $n = \sum_{i=1}^{r} n_i$，组平均值 $\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$，总平均值 $\bar{x} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij} = \frac{1}{n}\sum_{i=1}^{r} n_i\bar{x}_i$，则有平方和分解式

$$
\begin{aligned}
Q &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 + \sum_{i=1}^{r} n_i(\bar{x}_i - \bar{x})^2 = Q_1 + Q_2
\end{aligned}
$$

其中：Q 称为总离差平方和，简称总平方和，它反映全部数据之间的差异；Q_1 称为误差平方和或组内平方和，反映了随机误差的影响；Q_2 称为组间平方和或因素 A 的平方和，反映了各总体的样本平均值之间的差异，在一定程度上反映了 μ_i 间的差异程度，因而通过 Q_2 与 Q_1 的相对大小可以反映 H_0 是否成立。

若 Q_2 显著地大于 Q_1，说明 $\bar{x}$ 间的差异显著地大于随机误差，那么 H_0 可能不成立。这种比较方差大小以判断原假设是否成立的方法正是方差分析名称的由来。那么比值$\frac{Q_2}{Q_1}$大到什么程度可以否定原假设呢？从数理统计理论可知

$$
F = \left.\frac{Q_2/r-1}{Q_1/n-r}\right|_{H_0\text{成立}} \sim F(r-1,\ n-r)
$$

以 F 做为检验统计量，给定显著性水平 α，当 $P\{F(r-1,\ n-r)>F\}<\alpha$ 时，则拒绝 H_0，也就是说有 $1-\alpha$ 的把握认为因素对指标有显著影响，即 μ_i 间的差异是显著的。

通常，在进行方差分析时，要列出方差分析表 7.2.2。

表 7.2.2 方差分析表

方差来源	平方和	自由度	均方	F 值	显著性
因素	Q_2	$r-1$	$S_2^2=Q_2/(r-1)$	S_2^2/S_1^2	
误差	Q_1	$n-r$	$S_1^2=Q_1/(n-r)$		
总和	Q	$n-1$			

从计算角度讲,目前都采用现成的统计软件如 SPSS 等。对照第 1 节实例 2(见 SPSS 数据文件 ch7. 实例 2),我们采用 SPSS 软件可得到如下方差分析表。

表 7.2.3 ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2668.167	2	1334.083	10.930	.004
Within Groups	1098.500	9	122.056		
Total	3766.667	11			

因为 $P\{F(2, 9)>10.93\}=0.004\ll 0.05$,所以拒绝 H_0,即认为广告内容不同对销售量的影响是显著的。

2. 多重检验

如果 F 检验的结论是拒绝原假设,则表明从现在掌握的数据看,我们有理由认为因素 A 的 r 个水平效应有显著差异,也就是说,各总体均值不完全相同。为了寻求最优水平，我们还需进一步分析哪一些水平间差异是显著的,哪一些是不显著的。这类要同时比较多个水平之间指标差异是否显著的问题就是多重比较问题。因为这些水平之间两两作 T 检验,会使犯第一类错误的概率增大，所以 Fisher 建议,若原来显著性水平为 α,共作 N 次试验,则把显著性水平改为 α/N,仍用多个 T 检验。下面我们介绍两种方法，第一种是 Tukey 提出的 T 法，另一种是 Scheffe 提出的 S 法。

T 法的检验统计量和显著性水平是 α 的临界值分别为

$$d_{ij}=|\bar{x}_i-\bar{x}_j|,\ d_T=q_\alpha(r,\ n-r)\sqrt{\frac{Q_1}{m(n-r)}}$$

其中 $m=n_i(i=1, 2, \cdots, r)$, $q_\alpha(r, n-r)$可查表得到。当 $d_{ij}>d_T$ 时,认为第 i 水平与第 j 水平有显著差异。

T 法仅适用于不同水平下重复试验次数相等的情况，当各水平下试验次数不等时须用 S 法。设 $L=\sum_{i=1}^{r}c_i\mu_i$,其中 $c_i(i=1, 2, \cdots, r)$ 不全为 0,且 $\sum_{i=1}^{r}c_i=0$。$\hat{L}=\sum_{i=1}^{r}c_i\bar{x}_i$,则当

$$|\hat{L}| > d_s = \sqrt{(r-1)F_\alpha(r-1, n-r)\frac{Q_1}{n-r}\sum_{i=1}^{r}\frac{c_i^2}{n_1}}$$

时，认为假设 H_0： $L=0$ 不成立。此检验也可用在两两比较上，只要取 $|\hat{L}|=d_{ij}$ 即可。

我们还可以对每一对 μ_i 和 μ_j 之间的差异程度作出估计，这就要对效应之差 $\mu_i-\mu_j$ 作区间估计。由数理统计理论知识，对固定的 i，j，$\mu_i-\mu_j$ 的置信系数 $1-\alpha$ 的置信区间由最小显著差异法（LSD）可得

$$\left(\bar{x}_i-\bar{x}_j-\sqrt{\left(\frac{1}{n_i}+\frac{1}{n_j}\right)S_1^2}\ t_{\frac{\alpha}{2}}(n-r),\ \bar{x}_i-\bar{x}_j+\sqrt{\left(\frac{1}{n_i}+\frac{1}{n_j}\right)S_1^2}\ t_{\frac{\alpha}{2}}(n-r)\right)$$

由修正最小显著差异法（Bonferroni）可得

$$\left(\bar{x}_i-\bar{x}_j-\sqrt{\left(\frac{1}{n_i}+\frac{1}{n_j}\right)S_1^2}\ t_{\frac{\alpha}{2m}}(n-r),\ \bar{x}_i-\bar{x}_j+\sqrt{\left(\frac{1}{n_i}+\frac{1}{n_j}\right)S_1^2}\ t_{\frac{\alpha}{2m}}(n-r)\right)$$

其中 m 为区间的个数。如果这个区间包含零，则表明我们可以以概率 $1-\alpha$ 断言 μ_i 和 μ_j 没有显著差异；如果整个区间落在零的左边，则我们以概率 $1-\alpha$ 断言 μ_i 小于 μ_j；相反，如果整个区间落在零的右边，则我们以概率 $1-\alpha$ 断言 μ_i 大于 μ_j。

对于第 1 节实例 2（见 SPSS 数据文件 ch7。实例 2），我们采用 SPSS 软件可得如下结果。

表 7.2.4 Descriptives

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
1	4	173.50	9.33	4.66	158.66	188.34
2	4	187.75	8.18	4.09	174.73	200.77
3	4	209.75	14.57	7.28	186.57	232.93
Total	12	190.33	18.50	5.34	178.58	202.09

表 7.2.5 Test of Homogeneity of Variances

Levene Statistic	df_1	df_2	Sig.
1.091	2	9	.377

表 7.2.6 **Multiple Comparisons**

Dependent Variable：销售量

	(I)	(J)	Mean Difference (I−J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	1	2	−14.25	7.812	.101	−31.92	3.42
		3	−36.25	7.812	.001	−53.92	−18.58
	2	1	14.25	7.812	.101	−3.42	31.92
		3	−22.00	7.812	.020	−39.67	−4.33
	3	1	36.25	7.812	.001	18.58	53.92
		2	22.00	7.812	.020	4.33	39.67
Bonferroni	1	2	−14.25	7.812	.304	−37.17	8.67
		3	−36.25	7.812	.004	−59.17	−13.33
	2	1	14.25	7.812	.304	−8.67	37.17
		3	−22.00	7.812	.061	−44.92	.92
	3	1	36.25	7.812	.004	13.33	59.17
		2	22.00	7.812	.061	−.92	44.92

由以上结果可见，三个水平对应的三个总体的方差具有齐性(第二个表)。第三种广告引起的销售量最多(第一个表、第三个表)，第三水平为最优水平。在今后的广告宣传中，应多宣传噪音低的优良性，这对增加销量有好处。同时，应进一步进行工艺改革以降低噪音。

例 7.2.1 某种型号化油器的原中小喉管的结构油耗较大。为节约能源，设想了两种改进方案以降低油耗指标——比油耗，现对用各种结构的中小喉管制造的化油器分别测得一批数据，试问中小喉管的结构对比油耗的影响是否显著？并提出改进方案。

原始数据表

原结构比油耗 231.0 232.8 227.6 228.3 224.7 225.5 229.3 230.3

改进方案 1 比油耗 222.8 224.5 218.5 220.2

改进方案 2 比油耗 224.3 226.1 221.4 223.6

解 应用 SPSS 软件(见 SPSS 数据文件例 7.2.1)可得如下结果。

表 7.2.7 **Descriptives**

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
1	8	228.6875	2.7420	.9694	226.3952	230.9798
2	4	221.5000	2.6696	1.3348	217.2521	225.7479
3	4	223.8500	1.9434	.9717	220.7577	226.9423
Total	16	225.6813	4.0082	1.0020	223.5454	227.8171

表 7.2.8 **Multiple Comparisons**

Dependent Variable：比油耗

	(I)	(J)	Mean Difference (I−J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	1	2	7.1875	1.569	.001	3.7979	10.5771
		3	4.8375	1.569	.009	1.4479	8.2271
	2	1	−7.1875	1.569	.001	−10.5771	−3.7979
		3	−2.3500	1.812	.217	−6.2639	1.5639
	3	1	−4.8375	1.569	.009	−8.2271	−1.4479
		2	2.3500	1.812	.217	−1.5639	6.2639

表 7.2.9 **Test of Homogeneity of Variances**

Levene Statistic	df_1	df_2	Sig.
.575	2	13	.576

表 7.2.10 **ANOVA**

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	155.646	2	77.823	11.855	.001
Within Groups	85.339	13	6.565		
Total	240.984	15			

由此可知不同的中小喉管结构的比油耗有显著差异，从上图可以看出改进方案 1 的比油耗最小，采用这种结构有可能节省油耗。

例 7.2.2　有以下四种产品：A_1，国外同类产品；A_2，本厂产品；A_3，国内甲厂同类产品；A_4，国内乙厂同类产品。现从各厂产品中分别取 2,10,6,6 个产品做 300 小时连续磨损老化试验，得变化率数据为

A_1 变化率　12　14

A_2 变化率　20　18　19　17　15　16　13　18　22　17

A_3 变化率　26　19　26　28　23　25

A_4 变化率　24　25　18　22　27　24

假定各厂变化率服从等方差的正态分布。

(1) 试问四个厂的产品的变化率有否显著差异？

(2) 若有差异的话，请进一步检验：① 外国产品与本厂产品有无显著差异；② 外国产品与国内产品有无显著差异；③ 本厂产品与国内产品有无显著差异。

解　利用 SPSS 软件（见 SPSS 数据文件例 7.2.2）可得如下结果。

表 7.2.11　Descriptives 变化率

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
1	2	13.00	1.41	1.00	.29	25.71
2	10	17.50	2.55	.81	15.68	19.32
3	6	24.50	3.15	1.28	21.20	27.80
4	6	23.33	3.08	1.26	20.10	26.56
Total	24	20.33	4.68	.95	18.36	22.31

表 7.2.12　Test of Homogeneity of Variances 变化率

Levene Statistic	df_1	df_2	Sig.
.358	3	20	.784

表 7.2.13　ANOVA 变化率

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	346.000	3	115.333	14.661	.000
Within Groups	157.333	20	7.867		
Total	503.333	23			

表 7.2.14 Contrast Coefficients

Contrast	1	2	3	4
1	1	−1	0	0
2	−3	1	1	1
3	0	2	−1	−1

表 7.2.15 Contrast Tests

	Contrast	Value of Contrast	Std. Error	t	df	Sig. (2 - tailed)
Assume equal variances	1	−4.50	2.17	−2.071	20	.051
	2	26.33	6.23	4.227	20	.000
	3	−12.83	2.40	−5.343	20	.000
Does not assume equal variances	1	−4.50	1.28	−3.503	2.600	.049
	2	26.33	3.59	7.338	2.020	.018
	3	−12.83	2.41	−5.316	18.937	.000

表 7.2.16 Multiple Comparisons

Dependent Variable:变化率

	(I)	(J)	Mean Difference ($I-J$)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	1	2	−4.50	2.173	.051	−9.03	3.19E−02
		3	−11.50	2.290	.000	−16.28	−6.72
		4	−10.33	2.290	.000	−15.11	−5.56
	2	1	4.50	2.173	.051	−3.19E−02	9.03
		3	−7.00	1.448	.000	−10.02	−3.98
		4	−5.83	1.448	.001	−8.85	−2.81
	3	1	11.50	2.290	.000	6.72	16.28
		2	7.00	1.448	.000	3.98	10.02
		4	1.17	1.619	.480	−2.21	4.54
	4	1	10.33	2.290	.000	5.56	15.11

续表 7.2.16

	(I)	(J)	Mean Difference (I−J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
		2	5.83	1.448	.001	2.81	8.85
		3	−1.17	1.619	.480	−4.54	2.21

由以上结果可以看出，四个厂产品的变化率有显著差异。又进行对比检验得出，外国产品与本厂产品无显著差异，而外国产品与国内产品、本厂产品与国内产品均有显著差异。

7.3　应用 SPSS 进行单因素方差分析

1. 建立数据文件

启动 Windows 后，双击 SPSS 图标或单击“开始→程序→SPSS ”，SPSS 软件开始运行。启动 SPSS 后，出现的界面是数据编辑窗口，它的底部有两个标签：Data View（数据视窗）和 Variable View（变量视窗）。在 Variable View 中定义变量，输入变量名，定义变量类型：Numeric（数值型）或 String（字串型）；定义列宽：Width；定义小数位：Decimal；输入变量标签、变量值标签等。将每个变量定义之后，在 Data View 进行数据的输入与编辑。逐列输入数据或字串，从而在数据编辑窗建立了数据文件。

比如，为某职业病防治院对 31 名石棉矿工中的石棉肺患者、可疑患者和非患者进行了用力肺活量测定，数据如下表所示，问三组石棉矿工的用力肺活量有无差别？

表 7.3.1　三组石棉矿工的用力肺活量

石棉肺矿工	1.8	1.4	1.5	2.1	1.9	1.7	1.8	1.9	1.8	1.8	2.0
可疑患者	2.3	2.1	2.1	2.1	2.6	2.5	2.3	2.4	2.4		
非患者	2.9	3.2	2.7	2.8	2.7	3.0	3.4	3.0	3.4	3.3	3.5

在变量视窗中定义两个变量：(1)组别 g，数值型，取值 1，2，3，分别代表三组石

棉矿工的石棉肺矿工、可疑患者和非患者；(2)肺活量 x，数值型。在数据视窗中输入数据。数据输入的文件结构为表 7.3.2。

表 7.3.2

	x	g
1	1.8	1
2	2.3	2
3	2.9	3
4	1.4	1
5	2.1	2
6	3.2	3
7	1.5	1
8	2.1	2
9	2.7	3
10	2.1	1
11	2.1	2
12	2.8	3
13	1.9	1
14	2.6	2
15	2.7	3
16	1.7	1
17	2.5	2
18	3.0	3
19	1.8	1
20	2.3	2
21	3.4	3
22	1.9	1
23	2.4	2
24	3.0	3
25	1.8	1
26	2.4	2
27	3.4	3
28	1.8	1
29	3.3	3
30	2.0	1
31	3.5	3

2. 进行方差分析

单击"Analyze→Compare Mean→One→Way ANOVA"，在 One→Way ANOVA 对话框中，分别指定方差分析的指标(Dependent List)和因素(Factor)，单击 OK 即输出单因素方差分析结果如表 7.3.3。

表 7.3.3　ANOVA 肺活量

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups					
Within Groups					
Total					

若 Sig. 值很小，则表明：三组石棉矿工的用力肺活量有差异，但并不表明任何两两组间矿工的用力肺活量均有差异。如要知道具体哪两组均数间有差异，还需要使用 One→Way ANOVA 的选择项进行多重比较。

3. 多重比较

在 One→Way ANOVA 对话框中，单击 Contrasts（指定一种要用 t 检验来检验的 Priori 对比）或单击 Post Hoc（指定一种多重比较检验，当方差齐性时有 14 种，当方差非齐性时有 4 种）。选择好后，单击 Continue（继续），返回 One→Way ANOVA 对话框。再单击 Option，打开选择输出统计量的对话框，单击 Descriptive，计算并输出：例数、均值、标准差、标准误、均数的 95% 置信区间、最小值和最大值。单击 Homogeneity-of-Variance，方差齐性检验，然后单击 Continue，返回 One→Way ANOVA 后，再单击 OK 即出现需要的一系列表格。

停下来想一想，在现实生活中是不是有许多可以用方差分析方法解决的问题。如根据美国 1986 年有关 48 个大陆州的综合犯罪报告，将 48 个州分成 7 个地区（表 7.3.4），研究这 7 个地区的犯罪率是否不同。

又如失业率是否因为地区的不同而不同；有一群老人分别住在 5 个疗养院，每个疗养院中有 20 人，给他们服用抗抑郁剂，试判断老人服药的剂量与所住的疗养院是否有关；按照一定标准，给各个专业打分，然后对各个二级学院进行比较；由受过训练的品尝家对几种同类食品进行品尝打分，然后进行比较；等等。希望读者能动脑筋用学习的知识解决一些实际问题。

表 7.3.4

州序号	犯罪率（每十万人）	地区	州序号	犯罪率（每十万人）	地区
1	147	新英格兰	25	164	南方
2	140		26	476	
3	149		27	675	
4	557		28	588	
5	226		29	1036	
6	426		30	334	
7	989	中大西洋	31	540	
8	572		32	558	
9	359		33	274	
10	423	中西部	34	395	
11	308		35	758	
12	800		36	436	西南方
13	804		37	659	
14	258		38	658	
15	285		39	726	
16	235		40	293	落基山区
17	263		41	524	
18	578		42	157	
19	51		43	222	
20	125		44	267	
21	369		45	719	
22	427	南方	46	437	太平洋岸
23	833		47	550	
24	306		48	920	

7.4 双因素方差分析

在许多实际问题中，往往需要同时考察几个因素对指标的影响。比如在第 1 节实例 2 中，对机械销售量的影响除了考察新闻广告形式外，常常还考察机械的销售价格。同时研究两个因素对试验指标的影响，就是双因素方差分析问题。

由于存在两个因素的影响，就产生一个新问题，两因素对指标的影响是否正好是它们每个因素对指标影响的迭加？比如第 1 节中实例 2，是否会产生这样的情

况，分别使销售量达到最高的广告和价格(在不低于成本的情况下)结合起来，会使销售量增加的幅度超过(或低于)二者分别增加的幅度之和。这种各个因素的不同水平的搭配所产生的新的影响在统计上称为交互作用。各因素间是否存在交互作用是多因素方差分析新产生的问题。下面先介绍无交互作用的双因素方差分析技术。

双因素方差分析的数据结构为(表 7.4.1)。

表 7.4.1

	B_1	B_2	B_3	…	B_s
A_1	x_{11}	x_{12}	x_{13}	…	x_{1s}
A_2	x_{21}	x_{22}	x_{23}	…	x_{2s}
…	…	…	…	…	…
A_r	x_{r1}	x_{r2}	x_{r3}	…	x_{rs}

1. 数学模型

若因素 A 的第 i 种效应和因素 B 的第 j 种效应分别记作 α_i，β_i 试验误差记作 ε_{ij}，那么

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}\,(i = 1, 2, \cdots, r;\ j = 1, 2, \cdots, s),$$

其中

$$\sum_{i=1}^{r}\alpha_i = 0,\ \sum_{j=1}^{s}\beta_j = 0,\ \hat{\mu} = \bar{x},$$

$$\hat{\alpha}_i = \bar{x}_{i\cdot} - \bar{x}\ (i = 1, 2, \cdots, r),\ \hat{\beta}_j = \bar{x}_{\cdot j} - \bar{x}\ (j = 1, 2, \cdots, s)$$

并且假定 $\varepsilon_{ij} \sim N(0, \sigma^2)$。

2. 方差分析

判断因素 A 是否显著，等价于检验假设 H_{01}：$\alpha_1=\alpha_2=\cdots=\alpha_r=0$，而判断因素 B 是否显著等价于检验假设 H_{02}：$\beta_1=\beta_2=\cdots=\beta_s=0$。为了检验这些假设，同单因素方差分析类似，将总离差平方和 Q 进行分解

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{s}(x_{ij} - \bar{x})^2 = Q_1 + Q_2 + Q_3$$

其中 $Q_1 = s\sum(\bar{x}_{i.} - \bar{x})^2$，$Q_2 = r\sum(\bar{x}_{.j} - \bar{x})^2$，$Q_3 = \sum\sum(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} - \bar{x})^2$，选取检验统计量 $F_A = \frac{S_1^2}{S_3^2} = \frac{Q_1/(r-1)}{Q_3/(r-1)(s-1)}$，$F_B = \frac{Q_2/(s-1)}{Q_3/(r-1)(s-1)}$，分别对两个假设进行检验。

为方便起见，通常也列出如下方差分析表 7.4.2。

表 7.4.2 双因素方差分析表

方差来源	平方和	自由度	均方	F 值	显著性
A	Q_1	$r-1$	S_1^2	S_1^2/S_3^2	
B	Q_2	$s-1$	S_2^2	S_2^2/S_3^2	
误差	Q_3	$(r-1)(s-1)$	S_3^2		
总和	Q	$rs-1$			

关于计算问题，我们同样利用 SPSS 软件来实现。

例 7.4.1 为提高某种产品的合格率，考察原料用量和来源地对其是否有影响。原料来源地有三个：甲、乙、丙。原料用量有三种：现用量、增加 5%、增加 8%。每个水平组合各做一次试验，得到的数据如表 7.4.3 所示。

表 7.4.3

	现用量(1)	增加 5%(2)	增加 8%(3)
甲地(1)	59	70	66
乙地(2)	63	74	70
丙地(3)	61	66	71

试分析原料用量及来源地对产品合格率的影响是否显著。

解 设原料来源地为因素 A，三个地区为因素 A 的三个水平，第 i 个水平对合格率的特殊效应为 $\alpha_i(i=1, 2, 3)$；原料用量为因素 B，三种用料量为因素 B 的三个水平，第 j 个水平对合格率的特殊效应为 $\beta_j(j=1, 2, 3)$，则原假设为

$$H_{01}: \alpha_1 = \alpha_2 = \alpha_3 = 0 \qquad H_{02}: \beta_1 = \beta_2 = \beta_3 = 0$$

由 SPSS 软件（见 SPSS 数据文件例 7.4.1）可得如下结果见表 7.4.4。

表 7.4.4 **Tests of Between-Subjects Effects**

Dependent Variable:合格率

Source	Type III Sum of Squares	df	Mean Square	*F*	Sig.
Corrected Model	172.000	4	43.000	6.143	.053
Intercept	40000.000	1	40000.000	5714.286	.000
A	26.000	2	13.000	1.857	.269
B	146.000	2	73.000	10.429	.026
Error	28.000	4	7.000		
Total	40200.000	9			
Corrected Total	200.000	8			

这里显著性水平 $\alpha = 0.05$,从上表可见,不能拒绝 H_{01},但能拒绝 H_{02}。即根据现有数据资料,有95%的把握推断原料来源地对产品合格率影响不大,而原料用量对合格率有显著影响,见表7.4.5—表7.4.6。

表 7.4.5 **Estimates**

Dependent Variable:合格率

原料用量	Mean	Std. Error	95% Confidence Interval for Mean	
			Lower Bound	Upper Bound
1	61.000	1.528	56.759	65.241
2	70.000	1.528	65.759	74.241
3	69.000	1.528	64.759	73.241

表 7.4.6 **Multiple Comparisons**

Dependent Variable: 产品合格率

	(*I*)原料用量	(*J*)原料用量	Mean Difference (*I*−*J*)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	1	2	−9.00	2.160	.014	−15.00	−3.00
		3	−8.00	2.160	.021	−14.00	−2.00
	2	1	9.00	2.160	.014	3.00	15.00

续表 7.4.6

	(I)原料用量	(J)原料用量	Mean Difference (I－J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
		3	1.00	2.160	.667	－5.00	7.00
	3	1	8.00	2.160	.021	2.00	14.00
		2	－1.00	2.160	.667	－7.00	5.00
Bonferroni	1	2	－9.00	2.160	.042	－17.56	－.44
		3	－8.00	2.160	.062	－16.56	.56
	2	1	9.00	2.160	.042	.44	17.56
		3	1.00	2.160	1.000	－7.56	9.56
	3	1	8.00	2.160	.062	－.56	16.56
		2	－1.00	2.160	1.000	－9.56	7.56

由此可见，原料用量对产品合格率影响显著，B_2 水平为最优水平。原料来源地对产品合格率影响不显著，因而可考虑以方便且运费低为准则选择水平。比如乙地最方便且运输距离短，那么最优条件为 A_2B_2，即采用乙地原料并在原有用料量上增加 5%，这一方案为最佳。

例 7.4.2　根据下面资料(见表 7.4.7)分析不同地区和不同时间对农民家庭人均纯收入(单位:元)的影响。

表 7.4.7　农民家庭人均纯收入

时间＼地区	北京(1)	天津(2)	河北(3)	山西(4)	内蒙古(5)
1980 年(1)	290.46	277.92	175.78	155.78	181.32
1981 年(2)	350.67	297.77	204.41	179.53	225.14
1982 年(3)	432.63	326.12	235.73	227.18	273.03
1983 年(4)	519.48	411.69	298.07	275.78	294.20
1984 年(5)	664.16	504.64	345.00	355.50	336.12
1985 年(6)	775.08	564.55	385.23	358.32	360.41

请根据由 SPSS 软件得到的结果，做出统计分析。

解　利用 SPSS 可得以下结果(见表 7.4.8—7.4.10)。

表 7.4.8　Tests of Between－Subjects Effects

Dependent Variable：人均纯收入

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	559996.822(a)	9	62221.869	25.831	.000
Intercept	3526657.960	1	3526657.960	1464.080	.000
A	286626.752	5	57325.350	23.798	.000
B	273370.070	4	68342.517	28.372	.000
Error	48175.739	20	2408.787		
Total	4134830.521	30			
Corrected Total	608172.560	29			

a R Squared ＝ .921 (Adjusted R Squared ＝ .885)

由此可见，时间和地区两个因素对人均纯收入都有显著差异。下面仅对时间因素进行分析。

表 7.4.9　Estimates

Dependent Variable：人均纯收入

时间	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1980 年	217.052	21.949	171.267	262.837
1981 年	251.504	21.949	205.719	297.289
1982 年	298.938	21.949	253.153	344.723
1983 年	359.844	21.949	314.059	405.629
1984 年	441.084	21.949	395.299	486.869
1985 年	488.758	21.949	442.973	534.543

表 7.4.10　Multiple Comparisons

Dependent Variable：人均纯收入

LSD

(I) 时间	(J) 时间	Mean Difference ($I-J$)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1980 年	1981 年	−34.4520	31.04053	.280	−99.2014	30.2974
	1982 年	−81.8860(*)	31.04053	.016	−146.6354	−17.1366

续表 7.4.10

(I) 时间	(J) 时间	Mean Difference (I−J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
	1983 年	−142.7920(*)	31.04053	.000	−207.5414	−78.0426
	1984 年	−224.0320(*)	31.04053	.000	−288.7814	−159.2826
	1985 年	−271.7060(*)	31.04053	.000	−336.4554	−206.9566
1981 年	1980 年	34.4520	31.04053	.280	−30.2974	99.2014
	1982 年	−47.4340	31.04053	.142	−112.1834	17.3154
	1983 年	−108.3400(*)	31.04053	.002	−173.0894	−43.5906
	1984 年	−189.5800(*)	31.04053	.000	−254.3294	−124.8306
	1985 年	−237.2540(*)	31.04053	.000	−302.0034	−172.5046
1982 年	1980 年	81.8860(*)	31.04053	.016	17.1366	146.6354
	1981 年	47.4340	31.04053	.142	−17.3154	112.1834
	1983 年	−60.9060	31.04053	.064	−125.6554	3.8434
	1984 年	−142.1460(*)	31.04053	.000	−206.8954	−77.3966
	1985 年	−189.8200(*)	31.04053	.000	−254.5694	−125.0706
1983 年	1980 年	142.7920(*)	31.04053	.000	78.0426	207.5414
	1981 年	108.3400(*)	31.04053	.002	43.5906	173.0894
	1982 年	60.9060	31.04053	.064	−3.8434	125.6554
	1984 年	−81.2400(*)	31.04053	.017	−145.9894	−16.4906
	1985 年	−128.9140(*)	31.04053	.000	−193.6634	−64.1646
1984 年	1980 年	224.0320(*)	31.04053	.000	159.2826	288.7814
	1981 年	189.5800(*)	31.04053	.000	124.8306	254.3294
	1982 年	142.1460(*)	31.04053	.000	77.3966	206.8954
	1983 年	81.2400(*)	31.04053	.017	16.4906	145.9894
	1985 年	−47.6740	31.04053	.140	−112.4234	17.0754
1985 年	1980 年	271.7060(*)	31.04053	.000	206.9566	336.4554
	1981 年	237.2540(*)	31.04053	.000	172.5046	302.0034
	1982 年	189.8200(*)	31.04053	.000	125.0706	254.5694
	1983 年	128.9140(*)	31.04053	.000	64.1646	193.6634
	1984 年	47.6740	31.04053	.140	−17.0754	112.4234

Based on observed means

* The mean difference is significant at the .05 level

例 7.4.3 某酿造厂有三个化验员担任发酵粉的颗粒检验，今由三名化验员每天从该厂所生产的发酵粉中抽样一次，共抽 10 天，分别化验其中所含颗粒的百分率，化验结果如表 7.4.11 所示，问三名化验员的技术有无显著差异？这 10 天生产的发酵粉的颗粒百分率有无显著差异？

表 7.4.11

A \ B	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}
A_1	10.1	4.7	3.1	3.0	7.8	8.2	7.8	6.0	4.9	3.4
A_2	10.0	4.9	3.1	3.2	7.8	8.2	7.7	6.2	5.1	3.4
A_3	10.2	4.8	3.0	3.1	7.8	8.4	7.9	6.1	5.0	3.3

解 应用 SPSS 软件可得如下结果。

表 7.4.12 Tests of Between - Subjects Effects

Dependent Variable：百分率

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	164.396	11	14.945	1978.027	.000
Intercept	1058.508	1	1058.508	140096.647	.000
A	2.400E−02	2	1.200E−02	1.588	.232
B	164.372	9	18.264	2417.235	.000
Error	.136	18	7.556E−03		
Total	1223.040	30			
Corrected Total	164.532	29			

由此可知，三个化验员的化验技术没有显著差异，这 10 天生产的发酵粉的颗粒百分率有显著差异。

下面简单介绍有交互作用的双因素方差分析技术。

在前面的内容中已经知道，如果因素 A 和因素 B 没有交互作用，则只需要在各个组合水平下各做一次试验就可以进行方差分析。但是如果因素 A 和因素 B 有交互作用，这是必须在各个组合水平下做重复试验方可进行方差分析。下面通过一个例子来具体说明。

例 7.4.4 抗牵拉强度是硬橡胶的一项重要的性能指标，现试验考察下列两

个因素对该指标的影响。

A(硫化时间)：A_1(40 秒)，A_2(60 秒)。

B(催化剂种类)：B_1(甲种)，B_2(乙种)，B_3(丙种)。

六种组合水平下，各重复做了两次试验，测得数据(单位：kg/cm^2)如下表，试问因素 A、因素 B 对该指标的影响是否显著？

表 7.4.13

	B_1		B_2		B_3	
A_1	390	380	440	420	370	350
A_2	390	410	450	430	370	380

解 应用 SPSS 软件(见 SPSS 数据文件例 7.4.4)可得如下结果。

表 7.4.14 Tests of Between-Subjects Effects

Dependent Variable：强度

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	9866.667	5	1973.333	13.156	.003
Intercept	1904033.333	1	1904033.333	12693.556	.000
A	533.333	1	533.333	3.556	.108
B	9316.667	2	4658.333	31.056	.001
A. * B	16.667	2	8.333	.056	.946
Error	900.000	6	150.000		
Total	1914800.000	12			
Corrected Total	10766.667	11			

由此可见，因素 B 显著，而因素 A 和 A 与 B 交互作用都不显著。下面着重考察因素 B。

表 7.4.15 Estimates

Dependent Variable：强度

B	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	392.500	6.124	377.516	407.484
2	435.000	6.124	420.016	449.984
3	367.500	6.124	352.516	382.484

表 7.4.16　**Pairwise Comparisons**

Dependent Variable：强度

(I) B	(J) B	Mean Difference (I−J)	Std. Error	Sig.	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
1	2	−42.500	8.660	.008	−70.970	−14.030
	3	25.000	8.660	.083	−3.470	53.470
2	1	42.500	8.660	.008	14.030	70.970
	3	67.500	8.660	.001	39.030	95.970
3	1	−25.000	8.660	.083	−53.470	3.470
	2	−67.500	8.660	.001	−95.970	−39.030

Based on estimated marginal means，Adjustment for multiple comparisons：Bonferroni

表 7.4.17　***A* * *B***

Dependent Variable：强度

A	B	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
1	1	385.000	8.660	363.809	406.191
	2	430.000	8.660	408.809	451.191
	3	360.000	8.660	338.809	381.191
2	1	400.000	8.660	378.809	421.191
	2	440.000	8.660	418.809	461.191
	3	375.000	8.660	353.809	396.191

根据指标的实际意义，从以上结果可见，乙种催化剂、硫化时间 60 s 可使硬橡胶的抗牵拉强度达到最大。如果仅考虑显著性因素而不考虑非显著性因素，可以认为选乙种催化剂可使硬橡胶的抗牵拉强度达到最大，考虑成本问题，可选硫化时间 40 s。

7.5　应用 SPSS 进行双因素方差分析

1. 无交互作用情形

建立数据文件，包含三个变量：因素 1、因素 2、指标。

单击 Analyze→General Linear Model→Univariatel...，将指标变量移入 Dependent栏，两个因素都移入 Fixed Factor(s)栏；单击 Model...，选择 Custom，

将 Factors 栏中项全部移入 Model 栏中，选 Main effect，选 Include intercept in model，单击“Continue”；若对 Fixed Factor 进行多重比较，可单击 Post Hoc...，进行选项；单击“Plots”，确定 Horizontal Axis 和 Separate Lines，并击 Add，单击 Contrasts，指定每个 Factors，在 Contrast 下拉式菜单中选 Simple，击“Change”，并选 Reference 中的 Last 或 First，还可在 Options 的下拉式菜中进行选项，最后单击“OK”。

输出结果：

表 7.5.1 Tests of Between-Subjects Effects(各目标效应之间的检验)

Dependent Variable:

Source	Sum of Squares	df.	Mean Square	F	Sig.
Var1					
Var2					
Error					
Corrected Total					

2. 有交互作用情形

建立数据文件。

单击 Analyze→General Linear Model→Univariatel ...，在 Univariatel ... 对话框中，将指标变量移入 Dependent 栏，两个因素都移入 Fixed Factor(s)栏，单击“Options”按钮，在 GLM - General Factorial options 对话框中，将 Factor(s) and Factor 栏中的 OVERALL 放入 Display Means for. 栏，选择 Descriptive Statistics，Estimates of effect size 等，单击“Continue”按钮，单击“Plots”，确定 Horizontal Axis 和 Separate Lines，并击 Add，单击“Contrasts”，单击“OK”。

输出结果：

表 7.5.2 Tests of Between - Subjects Effects(各目标效应之间的检验)

Dependent Variable:

Source	Sum of Squares	df.	Mean Square	F	Sig.
Var1					
Var2					
Var1 * Var2					
Error					
Corrected Total					

等等。

例 7.5.1 某养猪场进行猪增重试验，选择四个品种(a)猪和三种饲料(b)，在 12 种搭配中，每种饲养一头，饲养三个月后测量增重(z)数据(表 7.5.3，见 SPSS 数据文件例 7.5.1)：

表 7.5.3

	z	a	b
1	51	1	1
2	53	1	2
3	52	1	3
4	56	2	1
5	57	2	2
6	58	2	3
7	45	3	1
8	49	3	2
9	47	3	3
10	42	4	1
11	44	4	2
12	43	4	3

试进行方差分析。

解 建立 SPSS 数据文件如表 7.5.3，并如图 7.5.1 依次点击。

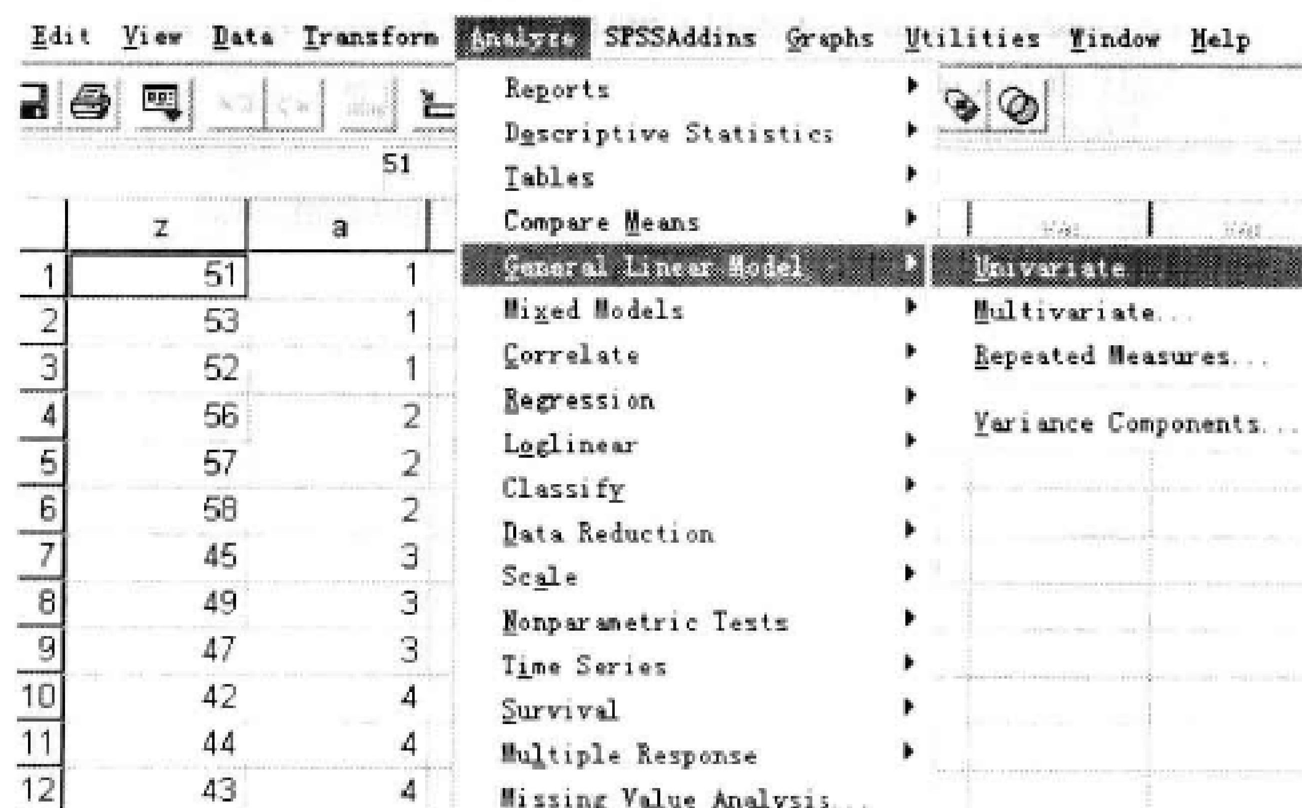


图 7.5.1

弹出对话框并填写相应选项(见图 7.5.2—图 7.5.4)。

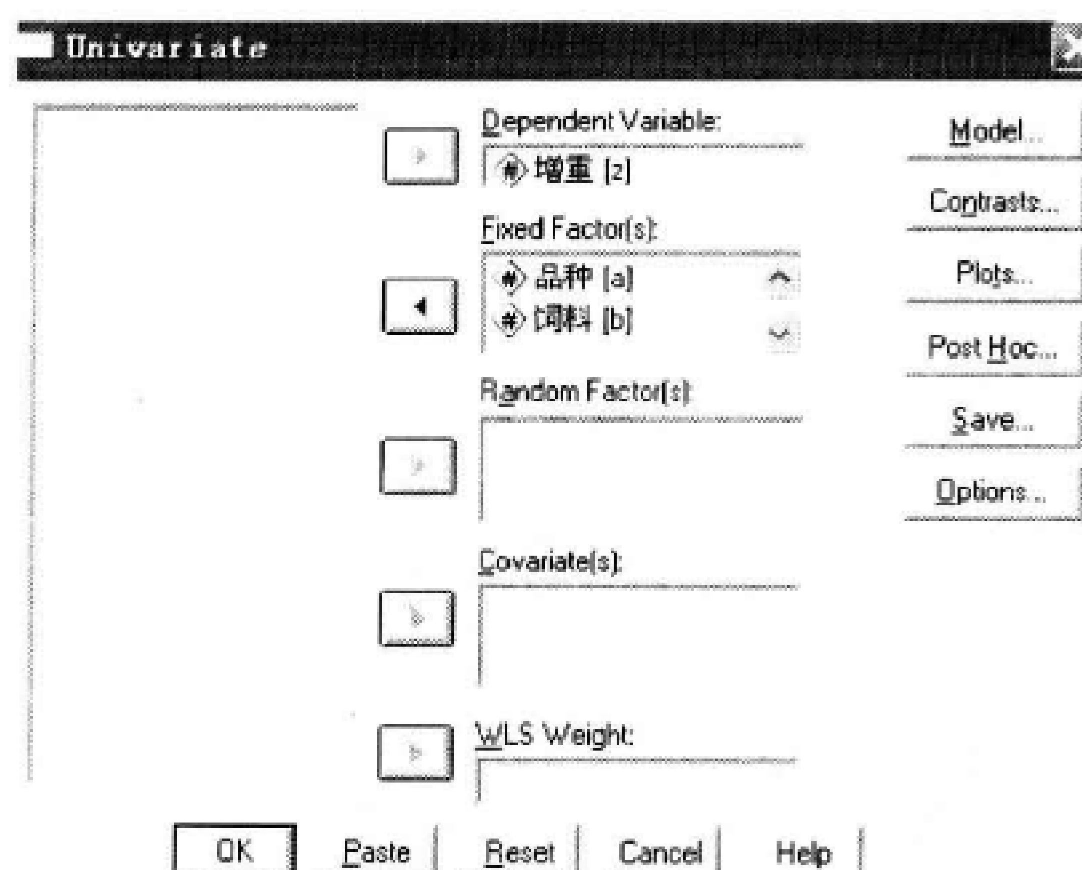


图 7.5.2

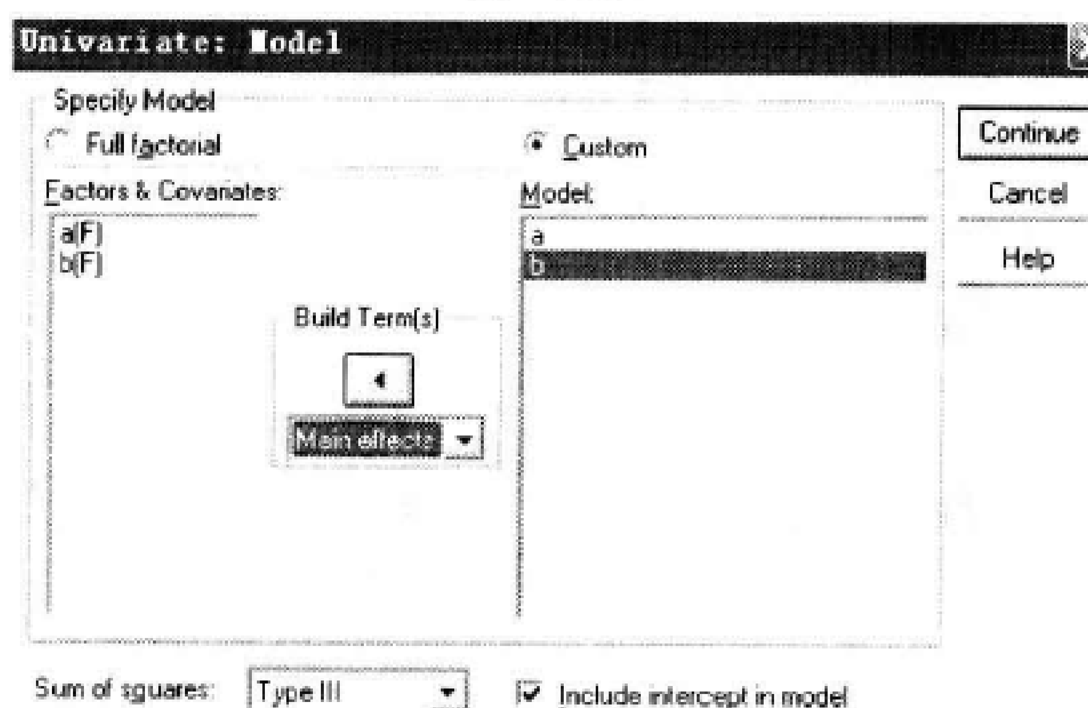


图 7.5.3

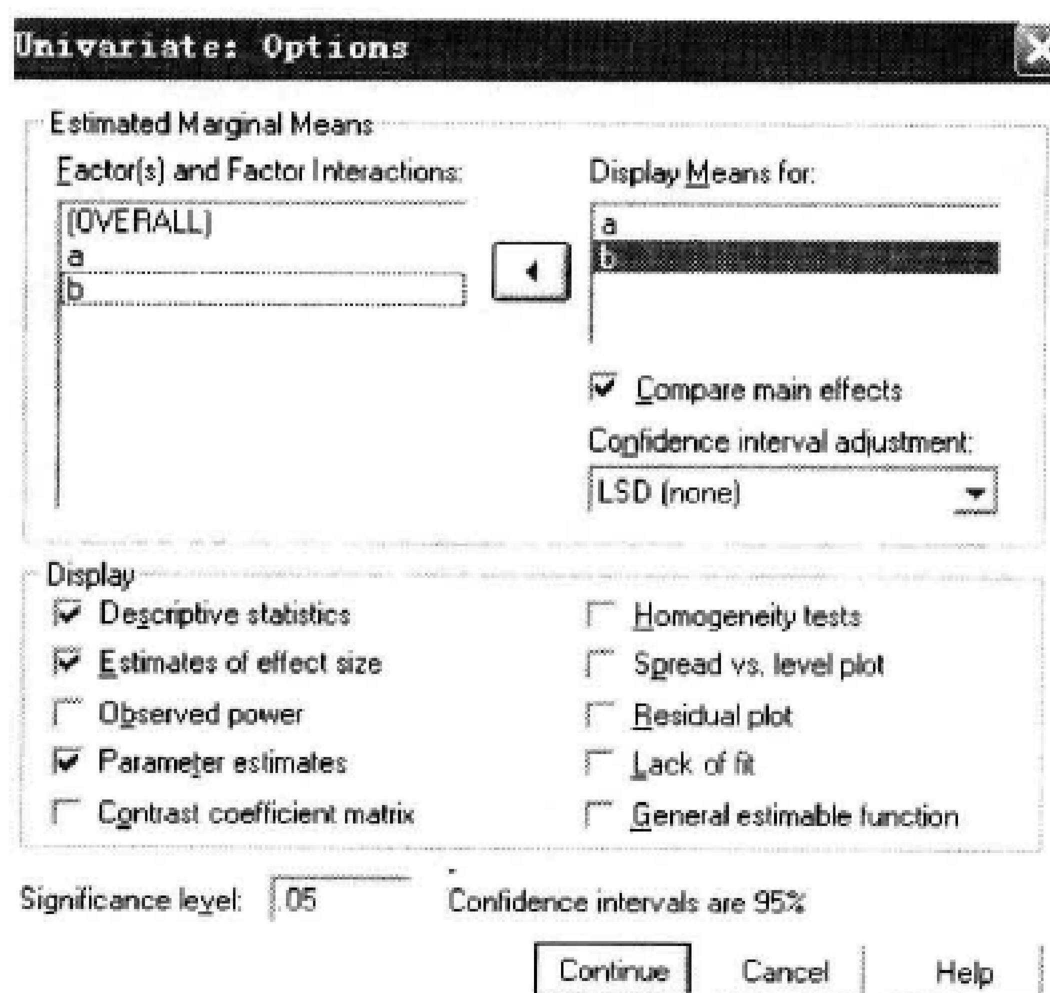


图 7.5.4

最后回到图 7.5.2 中点击"OK"，即输出以下结果。

Tests of Between-Subjects Effects

Dependent Variable：猪增重量

Source	Type III sum of Squares	df	Mean square	F	Sig.	Eta Squared
Corrected Model	342.750[a]	5	68.550	117.514	.000	.990
Intercept	29700.750	1	29700.750	50915.571	.000	1.000
A	332.250	3	110.750	189.857	.000	.990
B	10.500	2	5.250	9.000	.016	.750
Error	3.500	6	.583			
Total	30047.000	12				
Corrected Total	346.250	11				

a R Squared =. 990 (Adjusted R Squared =.981)

Univariate Tests

Dependent Variable：猪增重量

	Sum of Squares	df	Mean square	F	Sig.	Eta Squared
Contrast	332.250	3	110.750	189.857	.000	.990
Error	3.500	6	.583			

* The F tests the effect of 猪品种 This test is based on the linearly independent pairwise comparisons among the estimated marginal means

Univariate Tests

Dependent Variable：猪增重量

	Sum of Squares	df	Mean Qquare	F	Sig.	Eta Squared
Contrast	10.500	2	5.250	9.000	.016	.750
Error	3.500	6	.583			

* The F tests the effect of 饲料. This test is based on the linearly independent pairwise comparisons among the estimated marginal means

猪品种和猪饲料对猪的增重都有显著影响，且品种的影响比饲料的影响更大些。

例 7.5.2 用两种药物 A、B 治疗缺铁性贫血 12 例病人，服药一个月后，病人的红细胞增加数（百万/mm^3）数据如下：

表 7.5.4

		药物 B	
		不用	用
药物 A	不用	0.8，0.9，0.7	0.9，1.1，1.0
	用	1.3，1.2，1.1	2.1，2.2，2.0

试分析 A、B 药对红细胞增加的作用以及分析 A、B 药之间是否有交互作用？

解 应用 SPSS 软件（见 SPSS 数据文件例 7.5.2）可得：

表 7.5.5 Descriptive Statistics

Dependent Variable：红细胞增加数

药物 A	药物 B	Mean	Std. Deviation	N
1	1	.800	.100	3
	2	1.000	.100	3
	Total	.900	.141	6
2	1	1.200	1.000E−01	3
	2	2.100	.100	3
	Total	1.650	.501	6
Total	1	1.000	.237	6
	2	1.550	.609	6
	Total	1.275	.526	12

表 7.5.6 Tests of Between - Subjects Effects

Dependent Variable：红细胞增加数

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2.962	3	.987	98.750	.000
Intercept	19.507	1	19.507	1950.750	.000
A	1.688	1	1.688	168.750	.000
B	.907	1	.907	90.750	.000
A * B	.368	1	.368	36.750	.000
Error	8.000E−02	8	1.000E−02		
Total	22.550	12			
Corrected Total	3.042	11			

从偏差平方和的分解情况看，总偏差平方和 3.04 由两部分构成：可解释的偏差平方和 2.962 和由随机实验误差引起的残差平方和 0.080，而可解释的偏差平方和 2.962 由主效应 1.688+0.907=2.595 和交互效应 0.368 两部分偏差平方和构成。主效应药物 A、B 对红细胞增加 p 值<0.01，影响显著，交互效应的偏差平方和 0.368，自由度 1，均方 0.368，检验统计量 F 值为 36.750，p 值近似为 0.000，显然在 0.01 水平上有显著意义，说明交互效应对红细胞增加有显著影响。从各偏差平方和在总偏差平方和中所占的比例分析（等价从 F 值分析），药物 A 对红细胞增加数的贡献最大，疗效更显著一些，药物 B 次之，由于有交互作用存在，两种药物同时服用，效果更佳。

7.6 协方差分析

协方差分析是消除混杂因素的影响后进行的方差分析。比如，考虑药物对患者某个生化指标变化的影响，要比较实验组与对照组该指标的变化均值是否有显著性差异以确定药物的有效性，可能要考虑患者病程的长短、年龄以及原指标水平对疗效的影响。要消除这些因素的影响，考虑药物疗效，即比较实验组与对照组之间该生化指标变化量均值的差异显著性，才是科学的分析方法。只有在考虑了这些影响，在观测对象的选择上，使这些条件都一致时，才可以使用一般的方差分析方法。被消除的因素称为协变量，协变量是指一些与因变量、自变量可能都有关系的连续性变量，它们的存在可能会影响分析结果的正确性，从而不得不在分析中加以控制，这种控制了协变量的分析就是协方差分析。

例 7.6.1　镉作业工人按暴露于镉烟尘的年数分为大于等于 10 年和不足 10 年两组，两组工人的年龄未经控制（人随着年龄的增长，肺活量也会有所下降）测量了每个工人的肺活量，研究暴露于镉粉尘中的年龄与肺活量的关系，数据如下表（组别 1 代表大于等于 10 年，组别 2 代表不足 10 年），试进行协方差分析。

表 7.6.1

组别	年龄	肺活量	组别	年龄	肺活量
1	39	4.62	2	38	4.58
1	40	5.92	2	42	5.12
1	41	5.52	2	43	3.89
1	41	3.71	2	43	4.62
1	45	4.02	2	37	4.30

续表 7.6.1

组别	年龄	肺活量	组别	年龄	肺活量
1	49	5.09	2	50	2.70
1	52	2.07	2	50	3.50
1	47	4.31	2	45	3.06
1	61	2.70	2	48	4.06
1	65	3.03	2	51	4.51
1	58	2.73	2	46	4.66
1	59	3.67	2	58	2.88
2	43	4.61	2	38	3.64
2	39	4.73	2	38	5.09

解 利用 SPSS 软件(见 SPSS 数据文件例 7.6.1),选择 Analyze 、General Linear Model、Univariatel ...,在激活的对话框中,把肺活量放入 Dependent 栏,把分组放入 Fixed Factor(s)栏,把年龄放入 Covariate(s)栏,单击"Option"按钮,选择 Parameter estimates 和 Homogeneity tests,将分组放入 Display Means for 栏,单击"Continue"按钮,再单击"OK",即得到如下成果。

表 7.6.2 Descriptive Statistics

Dependent Variable:肺活量

分组	Mean	Std. Deviation	N
1	3.9492	1.1984	12
2	4.1219	.7677	16
Total	4.0479	.9592	28

表 7.6.3 Tests of Between-Subjects Effects

Dependent Variable:肺活量

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	11.085	2	5.543	10.073	.001
Intercept	41.936	1	41.936	76.216	.000
年龄	10.881	1	10.881	19.775	.000
分组	.542	1	.542	.985	.330

续表 7.6.3

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Error	13.755	25	.550		
Total	483.625	28			
Corrected Total	24.841	27			

表 7.6.4 Levene's Test of Equality of Error Variances

Dependent Variable：肺活量

F	df_1	df_2	Sig.
1.273	1	26	.270

* Tests the null hypothesis that the error variance of the dependent variable is equal across groups;Design：Intercept＋年龄＋分组

表 7.6.5 Parameter Estimates

Dependent Variable：肺活量

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	7.977	.886	8.998	.000	6.151	9.803
年龄	－8.700E－02	.020	－4.447	.000	－.127	－4.670E－02
[分组＝1]	.300	.303	.993	.330	－.323	.924
[分组＝2]	0					

This parameter is set to zero because it is redundant

表 7.6.6 Estimates

Dependent Variable：肺活量

分组	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	4.219	.223	3.761	4.678
2	3.919	.191	3.526	4.312

Evaluated at covariates appeared in the model：年龄 ＝ 46.64

由此可知，因变量为肺活量，因素变量为组别，协变量为年龄，进行接触镉粉尘

时间对肺活量影响的方差分析时消除受试者年龄引起的肺活量变化的影响。由方差分析的结果可以得出结论:肺活量的差异主要受试者年龄差异所致,与受试者接触镉粉尘的时间是否大于 10 年无关。

练习 7

1. 研究三种伤寒杆菌 j(因素、控制变量)对小白鼠存活天数 c(指标、观测变量)有无显著影响?

	c	j
1	2	1
2	4	1
3	3	1
4	2	1
5	4	1
6	7	1
7	7	1
8	2	1
9	5	1
10	4	1
11	5	2
12	6	2
13	8	2
14	5	2
15	10	2
16	7	2
17	12	2
18	6	2
19	6	2
20	7	3
21	11	3
22	6	3
23	6	3
24	7	3
25	9	3
26	5	3
27	10	3
28	6	3
29	3	3
30	10	3

2. 比较四种教学方法 g 的效果 x 是否存在明显差异,期末统考后,从这四个班各抽取 5 名学生的考试成绩:

	x	g
1	75	1
2	77	1
3	70	1
4	88	1
5	72	1
6	83	2
7	80	2
8	85	2
9	90	2
10	84	2
11	65	3
12	67	3
13	77	3
14	68	3
15	65	3
16	72	4
17	70	4
18	71	4
19	65	4
20	82	4

问:四种教学法的效果是否存在显著差异? 哪两种教学法间存在显著差异?

3. 对下列数据进行单因素方差分析。

1	2
27	22
31	27
31	25
29	23
30	26
27	27
28	23

计算 F 的观察值和检验的概值。比较概值和检验水平，确定是否应拒绝原假设。对数据进行独立样本 t 检验。比较 t 检验的概值和 F 检验的概值。结果有不同吗？（$\alpha=0.05$）

4. 新毕业的某专业学生在不同地区就业的起薪不同，这种情况看起来很合理。从三个地区随机地选取了一些用人单位，要求每个单位报告他们付给应届毕业生的起薪，收集数据如下。采用单因素方差分析对数据进行分析，在 $\alpha=0.05$ 的水平下进行假设检验，并讨论分析结果的意义。

地区 1	地区 2	地区 3
3050	4100	3550
3150	3950	3350
3000	3900	3500
3100	3800	3650
3150	3950	3600

5. 下面是 SPSS 的单因素方差分析输出结果，对这些结果进行分析，包括处理水平数目、样本容量、F 值、检验的统计显著性。

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1701	3	567	2.95	0.040
Within Groups	11728	61	192		
Total	13429	64			

6. 某化工公司的生意非常火。事实上生意火到公司的 5 个车间工人每周工

作时数超过 40 个小时。但是，管理者不清楚这 5 个车间的工人周工作时数是否存在差异。从每个车间随机抽取样本，利用 Excel 对数据进行分析，结果如下。解释本研究的设计，并判断在 $\alpha=0.05$ 的水平下均值之间是否存在显著差异。为什么有差异或者为什么没有差异？均值各为多少？研究结果对化工企业来说有什么意义？

	A	B	C	D	E	F	G
1	Anova:Singe Factor						
2							
3	SUMMARY						
4	Groups	Count	Sum	Average	Variance		
5	Plant 1	11	636.5577	57.87	63.5949		
6	Plant 2	12	601.7648	50.15	62.4813		
7	Plant 3	8	491.7352	61.47	47.4772		
8	Plant 4	5	246.0172	49.2	65.6072		
9	Plant 5	7	398.6368	56.95	140.354		
10							
11							
12	ANOVA						
13	urce of Variation	SS	df	MS	F	P-value	F crit
14	Between Groups	900.086	4	225.022	3.1	0.0266	2.62
15	Within Groups	2760.14	38	72.6352			
16							
17	Total	3660.22	42				

7. 某公司想对新销售人员进行不同的销售培训，为了比较培训课程的有效性，随机选择了三组销售人员，每组 5 人。一组接受 A 课程销售培训，一组接受 B 课程销售培训，另一组 C 没有参与任何训练。当前两组的训练课程结束时，收集训练后两个星期内的各组销售人员的销售记录如下：

三组销售人员销售业绩

A 课程	B 课程	C
2058	3339	2228
2176	2777	2578
3445	3020	1227
2517	2437	2044
944	3067	1681

判断在显著性水平为 0.1 的条件下是否有理由证明三组销售人员的销售水平有所

不同。

8. 一个大型汽车制造商希望了解四种品牌的轮胎（A，B，C 和 D）的平均里程数是否存在差异，因为该制造商想根据轮胎的耐用性来选择最好的供应商。供应商从每家公司选择了等级相同的轮胎，并在同等汽车上进行了检验。里程数如下：

A	B	C	D
31000	24000	30500	24500
25000	25500	28000	27000
28500	27000	32500	26000
29000	26500	28000	21000
32000	25000	31000	25500
27500	28000		26000
	27500		

在 $\alpha=0.05$ 下检验四种品牌的平均里程数是否存在显著差异，假定轮胎里程数服从正态分布。

9. 建筑工人委员会列出许多建筑工种，这些工种向工人支付的小时工资大体相当。其中包括砌砖、铁工和起重机驾驶。假定一个研究人员在全国范围内从这三个建筑工种中随机抽取了一个样本，询问工人的小时工资。假定调查获得的数据如下，这三个建筑工种的小时工资有显著差异吗？如果有显著差异，请利用多重比较程序判断哪些工种之间存在显著差异，$\alpha=0.05$。

砌砖	铁工	起重机驾驶
19.25	26.45	16.20
17.80	21.10	23.30
20.50	16.40	22.90
24.33	22.86	19.50
19.81	25.55	27.00
22.29	18.50	22.95
21.20		25.52
		21.20

10. 随着计算机技术的发展，对键盘的要求愈发严格。某公司要研究现有键盘设计是否影响打字员的速度，现随机选择 5 名打字员，每个打字员用三种键盘进行测试（每分钟打字个数）。

打字员	键盘 A	键盘 B	键盘 C
1	51	57	72
2	109	112	117
3	47	43	51
4	98	98	107
5	70	69	77

11. 某企业准备上市一种新型香水，需要进行市场调研。除香水气味外，经验表明香水包装与广告策略对销售量的增长也有很大影响。现用三个不同的广告策略和三种不同的包装对这种香水进行测试，每种组合采用两个不同的市场调查，调查结束后，数据如下：

各种促销手段下的销售量增长速度

包装设计 \ 广告策略	高雅	激情	流行
1	2.8	2.04	1.58
	2.73	1.33	1.26
2	3.29	1.5	1
	2.68	1.4	1.82
3	2.54	3.15	1.92
	2.59	2.88	1.33

根据这些数据，希望能找出最佳的促销模式。

12. 某商场有销售数据（见 SPSS 数据文件习题 7.12）。研究销售额是否受到促销方式、售后服务、奖金的影响？以及受到怎样的影响？

（1）进行不考虑交互效应和协变量（奖金）的方差分析；

（2）进行考虑交互作用但不考虑协变量的方差分析；

（3）进行方差分析。

13. 计算下面的方差分析表。确定检验的显著性（即计算检验的概值），并对各效应的假设进行判断。$\alpha=0.05$。

方差来源	SS	df	MS	F	Sig.
行	1.047	1			
列	3.844	3			
交互效应	0.773				
误差					
总和	12.632	23			

14. 考虑四台机器在三个不同的班次所生产出的阀门直径，质量控制员希望了解班次或机器对阀门平均直径是否有影响。按照机器和班次排列的数据如下表。此外，利用 SPSS 对数据进行双因素方差分析。该问题的假设是什么？对输出结果进行显著性检验。讨论所得到的结果。质量控制员从中能得到什么结论？

		阀门直径(cm)		
		班次		
		1	2	3
机器	1	6.56	6.38	6.29
		6.40	6.19	6.23
	2	6.54	6.26	6.19
		6.34	6.23	6.33
	3	6.58	6.22	6.26
		6.44	6.27	6.31
	4	6.36	6.29	6.21
		6.50	6.19	6.58

Tests of Between-Subjects Effects

Dependent Variable：直径(cm)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.233[a]	11	.021	1.604	.214
Intercept	964.568	1	964.568	73050.245	.000
JQ	.005	3	.002	.136	.937
BC	.197	2	.099	7.471	.008
JQ * BC	.030	6	.005	.383	.876
Error	.158	12	.013		
Total	964.959	24			
Corrected Total	.391	23			

a　R Squared = .595 (Adjusted R Squared = .224)

第 8 章　相关与回归分析

在实际问题的研究中，互有联系的变量间关系的紧密程度各不相同。一种极端的情况是一个(些)变量的变化能完全决定另一个(些)变量的变化，变量间完全表现为一种确定性关系，即函数关系；另一种普遍的情况是变量间有着密切的联系，但它们并没有密切到由一个(些)可以完全确定另一个(些)的程度，它们是一种非确定性关系，即统计关系。在现代统计学中关于统计关系的研究已成为统计学中两个重要的分支，它们叫相关分析和回归分析。在应用中，这两种统计分析方法经常相互结合和渗透，但它们研究的侧重点和应用面不同，使得它们的研究方法也大不相同。

变量的统计关系在现实中普遍存在，就个别例子而言，统计关系有不确定性或随机性，但大量观测后会发现，统计关系又具有一定的规律性，这种规律性的描写或者刻画就是相关分析要解决的问题。相关分析着重研究变量之间统计关系的密切程度，一般用相关系数来度量。

回归分析(regression analysis)是根据变量观测数据分析变量间关系的最常用的统计分析方法，其主要任务是根据变量的观测数据定量地建立所关注的变量和影响它变化的变量之间的数学关系式，检验影响变量的显著程度和比较它们的作用大小，进而用一组变量的变化解释和预测另一个变量的变化。通常把变量观测数据称为样本。如果数学关系式描写了一个变量与另一个变量之间的关系，则称其为一元回归分析；如果数学关系式描写了一个变量与另多个变量之间的关系，则称其为多元回归分析，并且称这一个变量是被影响变量即**因变量(dependent Variable)**，称这多个变量是影响变量，即**自变量(independent Variable)**。线性回归分析是描述变量间统计关系的一种最重要的统计模型技术，本章将着重介绍它的建模原理，模型参数的最小二乘估计，回归方程的有关检验和回归模型的应用。

8.1　两个变量的相关分析

在第 4 章中，两个随机变量 X, Y 的相关系数定义为

$$\rho=\frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X)}\ \sqrt{\operatorname{var}(Y)}}$$

对 X 的观测值 $x_1, x_2, \cdots, x_n$，对 Y 的观测值 $y_1, y_2, \cdots, y_n$，则可得到样本相关系数为

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

此相关系数也称为皮尔逊(Pearson)相关系数。皮尔逊相关系数只能描写两个随机变量之间线性相关程度的强弱，其取值范围是[−1, 1]，绝对值越接近 1，关系越密切，绝对值越接近 0，关系越疏远。其取值大于零，称为正相关；取值小于零，称为负相关；取值等于零，称为不相关。在理论上也可以证明，r 是 ρ 的相合估计。

在实际应用中，如何度量"绝对值接近 1"的接近程度呢？作统计分析有这样一个术语：在 α 水平下 X, Y 线性关系显著，即 X, Y 都服从正态分布，$\rho=0$ 时采用 t 检验来确定 r 的显著性。

记 $t(n-2)$ 为服从自由度为 $n-2$ 的 t 分布的随机变量，统计量 $t=\dfrac{r}{\sqrt{\dfrac{1-r^2}{n-1}}}$，$p=2P(t(n-2)>|t|)$。若 $|t|>t_{\frac{\alpha}{2}}(n-2)$(或 $p<\alpha$)，则表明 r 在统计上是显著的(ρ 与零存在显著差异)，即在 α 水平下 X, Y 线性关系显著；否则表明 r 在统计上是不显著的，即在 α 水平下 X, Y 线性关系不显著。

皮尔逊相关系数仅适用于度量定距变量(或定比变量)，下面介绍一种可度量定类变量或定序变量的相关系数，即斯皮尔曼(Spearman)等级相关系数。同时，它也可以度量非线性相关关系。斯皮尔曼等级相关系数定义为

$$r_s=1-\frac{6\sum_{i=1}^{n}(x_i-y_i)^2}{n(n^2-1)}$$

其中，x_i, y_i 分别是两个变量 X, Y 的观测值分别按大小(或按优劣等)排位的等级(称为秩)，n 为样本容量。与皮尔逊相关系数类似，其取值范围是[−1, 1]，绝对值越接近 1，关系越密切，绝对值越接近 0，关系越疏远。其取值大于零，称为正相关；取值小于零，称为负相关；取值等于零，称为不相关。

X, Y 是否存在显著的等级相关也需要进行检验。当 $n>20$ 时，可以利用以下统计量进行等级相关系数的显著性检验。

$$t=r_s\sqrt{\frac{n-2}{1-r_s^2}}$$

当 $|t|>t_{\frac{\alpha}{2}}(n-2)$ 或 $p=2P(t(n-2)>|t|)<\alpha$ 时，表明 X, Y 存在显著的等级相关。

皮尔逊相关系数和斯皮尔曼等级相关系数都由 SPSS 容易得到。

例 8.1.1　某公司下属 15 个分公司，它们的销售额 x（万元）、广告费 y（万元）、销售人员 z（个）数据如下表 8.1.1。

表 8.1.1

	bianhao	x	y	z
1	1	7800	21	19
2	2	8400	19	20
3	3	6100	18	20
4	4	5200	15	15
5	5	9700	21	21
6	6	8900	20	19
7	7	10000	22	22
8	8	9300	24	24
9	9	6500	15	15
10	10	7300	19	18
11	11	4800	13	12
12	12	4500	11	12
13	13	6700	18	18
14	14	7500	20	19
15	15	9500	15	25

试研究销售额 x（万元）、广告费 y（万元）、销售人员 z（个）之间的相关关系。

解　首先建立 SPSS 数据文件（见 SPSS 数据文件例 8.1.1），调用相关分析程序（图 8.1.1）。

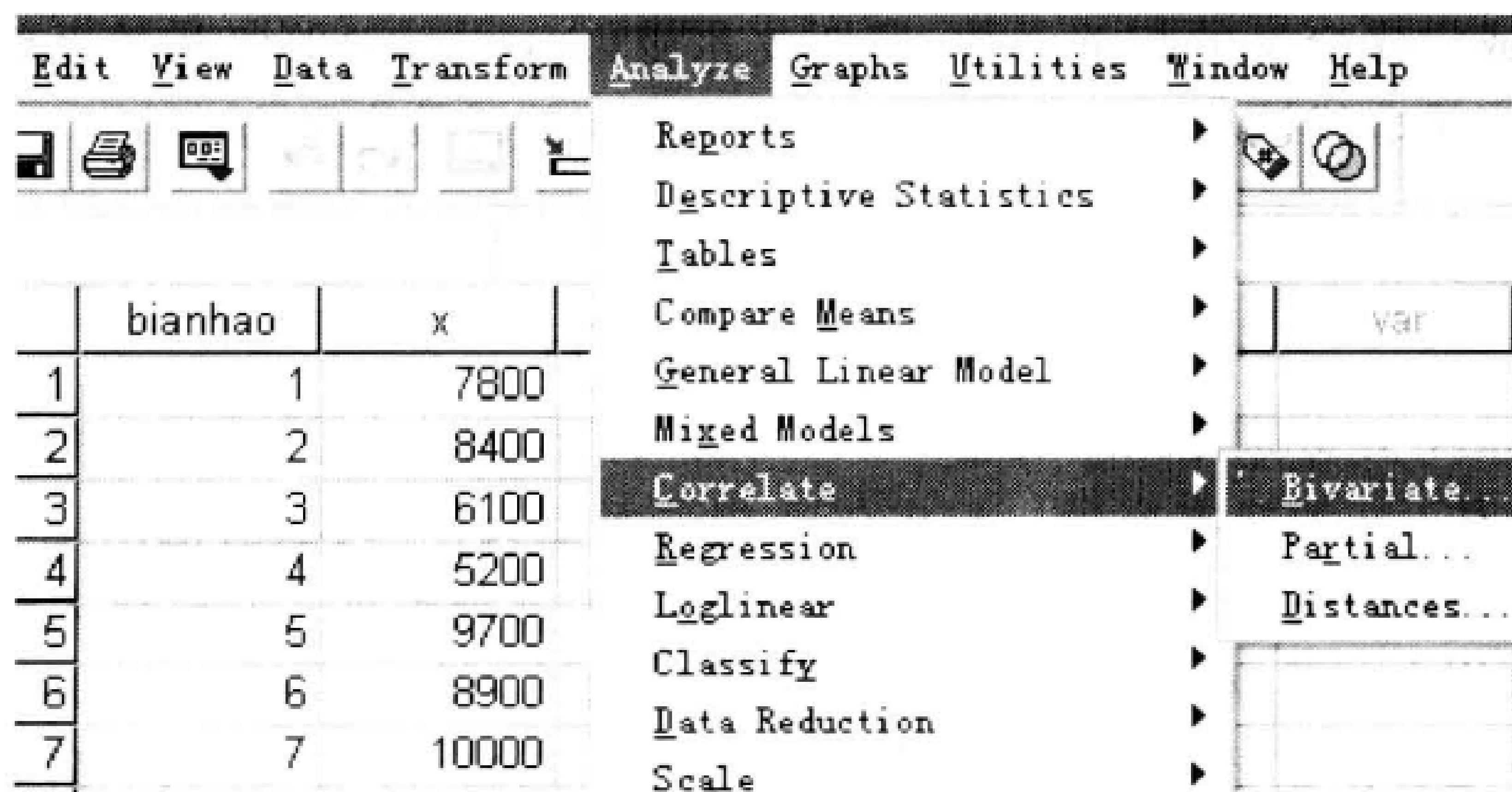


图 8.1.1

填写对话框（见图 8.1.2），点击"OK"，即得到表 8.1.2 和表 8.1.3 结果。

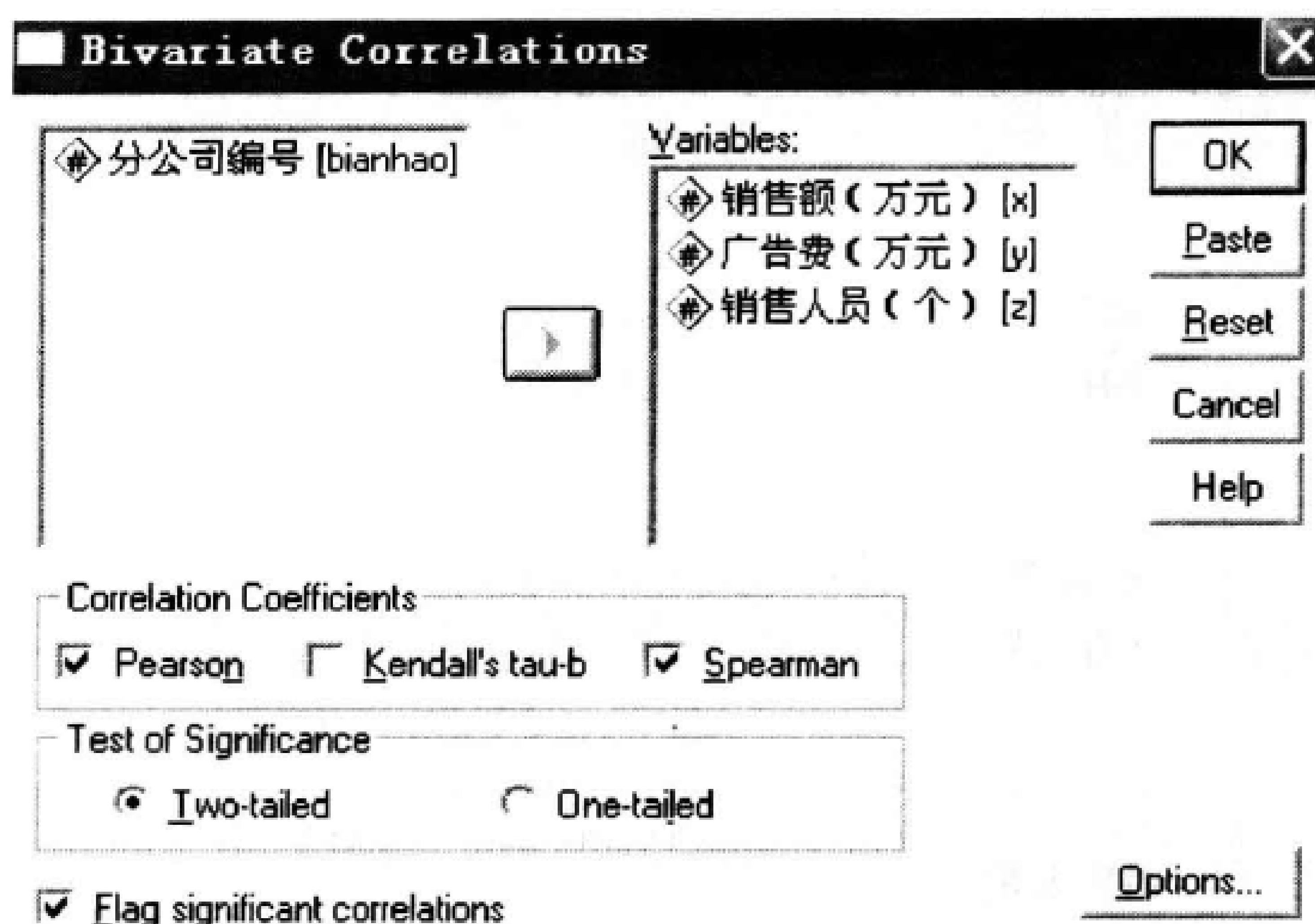


图 8.1.2

表 8.1.2　Correlations

		销售额 （万元）	广告费 （万元）	销售人员 （个）
销售 （万元）	Pearson Correlation Sig.（2-tailed） N	1 . 15	.766* .001 15	.884** .000 15
广告费 （万元）	Pearson Correlation Sig.（2-tailed） N	.766** .001 15	1 . 15	.718** .003 15
销售人员 （个）	Pearson Correlation Sig. (2-tailed) N	.884** .000 15	.718** .003 15	1 . 15

* Correlation is significant at the 0.01 level (2-tailed)

由此表可知：销售额 x 万元与广告费 y 万元的皮尔逊相关系数为 0.766，概值为 0.001；销售额 x 万元与销售人员 z 个的皮尔逊相关系数为 0.884，概值为1.206×10^{-5}；广告费 y 万元与销售人员 z 个的皮尔逊相关系数为 0.718，概值为0.003，在水平 0.01 下均有统计显著性。

表 8.1.3 **Correlations**

			销售额（万元）	广告费（万元）	销售人员（个）
Spearman's rho	销售额（万元）	Correlation Coefficient	1.000	.775**	.867**
		Sig. (2-tailed)	.	.001	.000
		N	15	15	15
	广告费（万元）	Correlation Coefficient	.775**	1.000	.661**
		Sig. (2-tailed)	.001	.	.007
		N	15	15	15
	销售人员（个）	Correlation Coefficient	.867**	.661**	1.000
		Sig. (2-tailed)	.000	.007	.
		N	15	15	15

* Correlation is siqnificant at the 0.01 level (2-tailed)

由此表可知：销售额 x 万元与广告费 y 万元的斯皮尔曼等级相关系数为 0.775，概值为 0.001；销售额 x 万元与销售人员 z 个的斯皮尔曼等级相关系数为 0.867，概值为 2.873×10^{-5}；广告费 y 万元与销售人员 z 个的斯皮尔曼等级相关系数为 0.661，概值为 0.007，在水平 0.01 下均有统计显著性。

在多变量的情况下，变量之间的相关关系很复杂。因此除了考虑两个变量的相关系数外，还要研究在控制其余变量情况下两个变量间的偏相关系数，研究一个变量与一组变量的线性相关性的复相关系数（或全相关系数），以及研究一组变量与另一组变量线性相关性的典型相关系数。偏相关系数和复相关系数在随后几节介绍，而典型相关系数请阅读相关书籍。

8.2 一元回归分析

在科学研究中，经常要涉及变量间的各种关系。比如，需求与价格、成本与产量、利润与价格、销售量等等。如果变量 y 与 x 间的关系可由方程 $y=a+bx$ 来描述，其中 a，b 均为常数，那么 y 与 x 间的关系是确定性的，即对于每个变量值 x 都只与某一个变量值 y 相对应。例如，某市场在 t 时刻黄瓜销量的数据如下（其中 q_t 表示 t 时刻销售黄瓜的数量，单位：斤，p_t 表示 t 时刻的销售价格，单位：元）：

表 8.2.1

p_t	q_t
2.5	1
2.0	3
1.5	5
1.0	7
0.5	9
0	11

这些结果可以用一个方程的形式概括为 $q_t=11-4p_t$ 这是一个确定性关系，因为对于每个价格，总有一个相应黄瓜售出量。如果对于任何已知的 x 值，变量 y 可以某个概率取某些特殊的值，则 x、y 之间的关系是随机的，例如

表 8.2.2

p_t	q_t	概率
2.5	0	0.25
	1	0.50
	2	0.25
2.0	2	0.25
	3	0.50
	4	0.25
…	…	…
0	10	0.25
	11	0.50
	12	0.25

这时，方程的形式为 $q_t=11-4p_t+\varepsilon_t$ 式中 ε_t 是一个具有分布

表 8.2.3

ε_t	概率
−1	0.25
0	0.50
1	0.25

的随机变量，称为随机扰动或随机误差项。由于随机扰动的存在，对于每一价格，有几个销售量与之对应，而每一销售量的发生都具有某个确定的概率，所以，后一种关系是随机的。

在回归分析中，我们研究变量的随机性关系。其中最简单的是两个变量之间的线性关系，其回归模型为

$$y_i = a + bx_i + \varepsilon_i$$

式中 y 称为因变量，x 称为自变量，ε 称为随机扰动，a，b 称为待估计的回归参数，下标 i 表示第 i 个观测值。

对于回归模型，我们假设

$$\varepsilon_i \sim N(0,\ \sigma^2),\ i = 1,\ 2,\ \cdots,\ n$$
$$E(\varepsilon_i\varepsilon_j) = 0,\ i \neq j$$

这时可得到

$$y_i \sim N(a + bx_i,\ \sigma^2)$$

于是，当 $x=0$ 时，y 的平均值为 a；当 x 取值发生变化时，y 的平均值变化 b。例如，y 表示总消费，x 表示总收入，则 a 表示消费的维持生活水平，b 则表示边际消费倾向。如果给出 a 和 b 的估计量分别为 $\hat{a}$，$\hat{b}$，则经验回归方程为

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

一般地，y_i 的值分布在 $\hat{y}_i$ 为中心的周围，它们之间的差称为残差，用 e_i 表示，即 $e_i = y_i - \hat{y}_i$，由于 $\hat{a}$，$\hat{b}$ 不同于 a，b 的真值，e_i 与 ε_i 是不相同的。实际上，残差 e_i 可视为扰动 ε_i 的“估计量”。

8.3 回归系数的最小二乘估计

设对 y 及 x 作 n 次观测得 n 组数据 $(y_i,\ x_i)$，$i=1,\ 2,\ \cdots,\ n$，应用计算机软件 Excel 做散点图，当散点呈直线趋势时，我们认为 y 与 x 的关系可用一元回归模型来描述。根据最小二乘原理，即让离差平方和达到最小的原则求解回归系数 a，b 的估计量 $\hat{a}$，$\hat{b}$，即

$$\min Q(a,b) = \min\sum_{i=1}^{n}(y_i - E(y_i))^2 = \min\sum_{i=1}^{n}(y_i - a - bx_i)^2$$

由于 $Q(a,\ b)$ 为 a，b 的二次非负函数，其最小值存在，根据微积分求极值的方法，a，b 应满足

$$\frac{\partial Q}{\partial a} = 0,\ \frac{\partial Q}{\partial b} = 0$$

容易解出

$$\hat{a} = \bar{y} - \hat{b}\bar{x},\quad \hat{b} = \frac{l_{xy}}{l_{xx}}$$

其中 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，$\bar{y}=\frac{1}{2}\sum_{i=1}^{n}y_i$，$l_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2$，$l_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$。

例 8.3.1　某市场连续 12 天卖出黄瓜的价格和数量的调查数据如下。

表 8.3.1

价格 x_i(元/斤)	销量 y_i(斤)
1.00	55
0.90	70
0.80	90
0.70	100
0.70	90
0.70	105
0.70	80
0.65	110
0.60	125
0.60	115
0.55	130
0.50	130

试求：黄瓜销量对价格的**回归方程(regression equation)**。

解　打开 SPSS，建立数据文件(见 SPSS 数据文件例 8.3.1)

表 8.3.2

	x	y
1	1.00	55
2	.90	70
3	.80	90
4	.70	100
5	.70	90
6	.70	105
7	.70	80
8	.65	110
9	.60	125
10	.60	115
11	.55	130
12	.50	130

调用线性回归程序：单击 Analyze→Regression→Linear，打开线性回归对话框。在左侧源变量栏中选择因变量 y(销量)进入 Dependent 栏中，选择自变量 x

(价格)进入 Independent 栏中。单击“Statistics”按钮,在对话框中可选择:回归系数的置信区间(confidence Intervals),描述性统计(descriptives),单击“Continue”→“OK”按钮,便得到程序“默认选择项”的结果。

表 8.3.3 Descriptive Statistics

	Mean	Std. Deviation	N
销量	100.00	23.932	12
价格	.7000	.14302	12

表 8.3.4 Correlations

		销量	价格
Pearson Correlation	销量	1.000	−.943
Sig. (1-tailed)	价格	−.943	1.000
	销量	.	.000
	价格	.000	.
N	销量	12	12
	价格	12	12

表 8.3.5 Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	210.444	12.571		16.741	.000
	价格	−157.778	17.624	−.943	−8.952	.000

a Dependent Variable: 销量

可得经验回归方程为

$$\hat{y}_i = 210.444 - 157.778 x_i$$

这就是我们要估计的需求函数。由于是线性函数,需求的价格弹性随价格的不同而变化,在平均价格($x_i = 0.7$)这一点上,价格弹性 η 的估计量为

$$\hat{\eta}_{x_i = x = 0.7} = -157.778 \times 0.007 = -1.104$$

这说明需求在这一点具有较小的弹性。

8.4　回归估计的统计推断

上一节，我们在最小二乘意义下推导了回归参数的估计式以及 SPSS 的实现，下面要进一步探讨经典正态线性回归模型中其估计式的其它特征以及统计推断。

1. 可以证明

(1)估计量 $\hat{a}$，$\hat{b}$ 分别是 a、b 的无偏估计量；

(2)$\text{var}(\hat{a})=\sigma^2(\frac{1}{n}+\frac{\overline{x}^2}{l_{xx}})$，$\text{var}(\hat{b})=\frac{\sigma^2}{l_{xx}}$；

(3)由于 $\hat{a}$，$\hat{b}$ 均为相互独立正态变量 y_1，y_2，…，y_n 的线性组合，根据正态分布的性质，它们也一定是正态的。

2. 从回归估计量(regression estimator)的方差可以看到

(1)扰动 ε_i 的方差 σ^2 越大，$\hat{a}$，$\hat{b}$ 的方差也越大，这意味着围绕总体回归直线的扰动离差越大，我们所估计的参数的离差也越大，如果所有扰动为零，那么关于回归参数的估计量就和它们的真值一致；

(2)自变量 x 的值越分散，$\hat{a}$，$\hat{b}$ 的方差越小，当 $x_1=x_2=\cdots=x_n$ 时，$\hat{a}$，$\hat{b}$ 的方差变成无穷大，这表明 x 的离差越大，l_{xx} 也越大，$\hat{a}$，$\hat{b}$ 就越接近于 a，b 的真值；

(3)当 $\overline{x}=0(l_{xx}\neq0)$ 时，$\hat{a}$ 的方差最小。

如上对自变量值的选择，通常是不可能的，因为我们只能在别人抽样结果的基础上来工作。

3. 总体方差 σ^2 的一个无偏估计量是

$$S^2=\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\frac{1}{n-2}\sum_{i=1}^{n}e_i^2$$

用 S^2 代替 σ^2，我们可以得到 $\hat{a}$，$\hat{b}$ 方差的无偏估计量分别是

$$S_{\hat{a}}^2=S^2(\frac{1}{n}+\frac{\overline{x}^2}{l_{xx}})，S_{\hat{b}}^2=\frac{S^2}{l_{xx}}$$

以后我们把这两个无偏估计量的算术平方根分别称为 a，b 的估计标准误差。

4. a 和 b 的区间估计

置信水平为 $1-\alpha$ 的区间估计是

$$(\hat{a}-t_{\frac{\alpha}{2},n-2}S_{\hat{a}},\ \hat{a}+t_{\frac{\alpha}{2},n-2}S_{\hat{a}})$$
$$(\hat{b}-t_{\frac{\alpha}{2},n-2}S_{\hat{b}},\ \hat{b}+t_{\frac{\alpha}{2},n-2}S_{\hat{b}})$$

5. $E(y_i)$的区间估计

因为总体回归直线为 $E(y_i)=a+bx_i$，而它的估计式为 $\hat{y}_i=\hat{a}+\hat{b}x_i$，又 $E(\hat{y}_i)=a+bx_i=E(y_i)$，所以 $\hat{y}_i=\hat{a}+\hat{b}x_i$ 是 $E(y_i)$的一个无偏估计量。为了对 $E(y_i)$进行区间估计，我们首先可求出 $\hat{y}_i=\hat{a}+\hat{b}x_i$ 的方差 $\sigma^2_{\hat{y}_i}=\sigma^2\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{l_{xx}}\right)$，又因$\hat{y}_i\sim N(E(\hat{y}_i),\ \sigma^2_{\hat{y}_i})$，并记 $S^2_{\hat{y}_i}=S^2\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{l_{xx}}\right)$，则$\frac{\hat{y}_i-E(\hat{y}_i)}{S_{\hat{y}_i}}\sim t(n-2)$，这样便可得 $E(y_i)$的置信水平为 $1-\alpha$ 的区间估计是

$$(\hat{y}_i-t_{\frac{\alpha}{2},n-2}S_{\hat{y}_i},\ \hat{y}_i+t_{\frac{\alpha}{2},n-2}S_{\hat{y}_i})$$

6. y 的样本变差的分解

易见 $$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

离差平方和　　回归平方和　　残差平方和

这样，y 的样本变差(SST)就分解为两部分，一部分是回归平方和(SSR)反映了回归自变量变差的贡献，另一部分残差平方和(SSE)反映了误差的影响。由于SST＝SSR＋SSE，记 $R^2=\mathrm{SSR}/\mathrm{SST}$，则 R^2 度量了经验回归方程对观测数据的拟合程度，称之为判定系数。显然 $0\leqslant R^2\leqslant 1$，$R^2$是因变量与自变量相关系数的平方。因此，它的值越大，表明因变量与自变量之间的相关性就越强。

7. 回归方程的显著性检验

前面我们解决了简单回归模型的参数估计问题，并给出了一元线性经验回归方程，然而所得回归方程代表性究竟如何？即自变量与因变量之间是否真正具有线性关系？所以必须进行**显著性检验(significance test)**，即评价所建立的回归方程是否有显著意义。如果回归方程无显著意义，那么其自变量前的系数可以取值为零，反之就不能取值零。所以，这时原假设和备择假设可以这样来提出

$$H_0: b=0 \quad H_1: b\neq 0$$

要检验 H_0，我们在经典正态回归模型下，构成检验统计量 $Z=\frac{\hat{b}-b}{S_{\hat{b}}}$。在 H_0 下，$Z=\frac{\hat{b}}{S_{\hat{b}}}\sim t(n-2)$。对于给定的显著性水平 α，当 $P\{t(n-2)>|Z|\}<\alpha$ 时，就拒绝

H_0，认为回归方程有显著意义。另外，注意到在因变量的样本变差分解中，SST＝SSR＋SSE＝$\hat{b}^2 l_{xx}$＋SSE，若 H_0 成立，则 SSR＝0，SST＝SSE，这样因变量的变差全部由于随机误差所致，我们可构成检验统计量 $F=\frac{\mathrm{SSR}/1}{\mathrm{SSE}/(n-2)}$，当 $P\{F(1, n-2)>F\}<\alpha$ 时，就拒绝 H_0，认为回归方程有显著意义。如上两种检验结果是相同的，而后一种检验可用于多个自变量的情况。

8. 回归分析的表述

我们从一组样本数据进行回归系数的估计，得到经验回归方程，因为还要进行区间估计、显著性检验，所以必须求出回归估计量的标准误 $S_{\hat{a}}$，$S_{\hat{b}}$，以及判定系数 R^2，通常可写成表达式：

$$\hat{y}_i=\hat{a}+\hat{b}x_i, R^2= \cdots$$
$$(\quad)\quad(\quad)$$

其中括号内填写相应的 t－检验显著性概率值。这样就较全面地表述了样本回归估计式。

例 8.4.1　对于例 8.3.1 黄瓜需求问题，我们应用 SPSS（见 SPSS 数据文件例 8.3.1）可得下面成果：

表 8.4.1　Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.943[a]	.889	.878	8.360

a　Predictors：(Constant)，价格

表 8.4.2　ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5601.111	1	5601.111	80.143	.000[a]
	Residual	698.889	10	69.889		
	Total	6300.000	11			

a　Predictors：(Constant)，价格

b　Dependent Variable：销量

表 8.4.3 Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	Constant	210.444	12.571		16.741	.000	182.435	238.454
	价格	−157.778	17.624	−.943	−8.952	.000	−197.047	−118.508

a Dependent Variable：销量

由此可知：

(1) $R^2 = 0.889$，$S = 8.360 = (698.889/10)^{0.5}$，平方和分解 SST$= 6300 = 5601.111 + 698.889 =$ SSR$+$SSE，$S_{\hat{a}} = 12.571$，$S_{\hat{b}} = 17.624$，这表明因变量的样本变差的 88.9% 是拟合所产生的，即由 $\hat{y}$ 对 y 的拟合。R^2 也表明了样本回归直线对观测值的拟合相当好；

(2)回归方程的显著性检验，从 t -检验和 F -检验均有显著性概率值 $p = 4.34E-06 \ll 0.05$，所以认为回归方程是显著的；

(3)a 的 95% 置信区间是 (182.435, 238.454)，b 的 95% 置信区间是 (−197.047, −118.508)

(4)最后可写出经验回归方程

$$\hat{y}_i = 210.444 - 157.778x_i$$
$$(1.21E-08)\ (4.34E-06)$$

下面我们求 $E(y_i)$ 的 95% 置信区间。

在回归分析主对话框中，单击保存新变量按钮“Save”，将选择的新变量保存在数据文件中，在对话框中可选择：未标准化预测值(Unstandardized)，对平均、个体的预测区间(Prediction Individual)等，即得到：

表 8.4.4

	x	y	pre_1	lmci_1	umci_1	lici_1	uici_1
1	1.00	55	52.66667	39.71667	65.61666	29.98026	75.35307
2	.90	70	68.44444	58.92615	77.96274	47.52631	89.36258
3	.80	90	84.22222	77.56376	90.88069	64.44077	104.00367
4	.70	100	100.00000	94.62281	105.37719	80.61225	119.38775
5	.70	90	100.00000	94.62281	105.37719	80.61225	119.38775
6	.70	105	100.00000	94.62281	105.37719	80.61225	119.38775
7	.70	80	100.00000	94.62281	105.37719	80.61225	119.38775
8	.65	110	107.88889	102.16443	113.61335	88.40197	127.37581
9	.60	125	115.77778	109.11931	122.43624	95.99633	135.55923
10	.60	115	115.77778	109.11931	122.43624	95.99633	135.55923
11	.55	130	123.66667	115.69100	131.64233	103.40385	143.92949
12	.50	130	131.55556	122.03726	141.07385	110.63742	152.47369

从表 8.4.4 中,Lmci-1 是 $E(y_i)$的置信下限,Uici-1 是 $E(y_i)$的置信上限。

8.5 预 测

我们建立了回归方程后,其重要应用之一就是用来进行预测。

1. 预测值(prediction value)

已知自变量 x 的值 x_0,预测因变量 y 的相应值 y_0,由于 $y_0=a+bx_0+\varepsilon_0$ 是一个散布在总体回归直线周围对应于 x_0 点的随机变量,所以实验前我们还不能知道它的取值,但可得到其预测值 $\hat{y}_0=\hat{a}+\hat{b}x_0$。

例 8.5.1 对于例 8.3.1,预测当价格分别为 1.2,1.1,0.85,0.75,0.45 时,黄瓜的销量情况。

解 打开 SPSS 建立的数据文件,将新价格填入数据文件中(见 SPSS 数据文件例 8.5.1),并在主对话框的保存对话框中选"未标准化预测"和"个体区间预测",运行即得:

表 8.5.1

	x	y	pre_1	lici_1	uici_1
1	1.00	55	52.66667	29.98026	75.35307
2	.90	70	68.44444	47.52631	89.36258
3	.80	90	84.22222	64.44077	104.00367
4	.70	100	100.00000	80.61225	119.38775
5	.70	90	100.00000	80.61225	119.38775
6	.70	105	100.00000	80.61225	119.38775
7	.70	80	100.00000	80.61225	119.38775
8	.65	110	107.88889	88.40197	127.37581
9	.60	125	115.77778	95.99633	135.55923
10	.60	115	115.77778	95.99633	135.55923
11	.55	130	123.66667	103.40385	143.92949
12	.50	130	131.55556	110.63742	152.47369
13	1.20	.	21.11111	-6.48250	48.70473
14	1.10	.	36.88889	11.93655	61.84123
15	.85	.	76.33333	56.07051	96.59615
16	.75	.	92.11111	72.62419	111.59803
17	.45	.	139.44444	117.71277	161.17611

每一价格对应的黄瓜销量分别为:

表 8.5.2 个体预测值

x	$\hat{y}$
1.20	21.11
1.10	36.88
0.85	76.33
0.75	92.11
0.45	139.44

实际值 y_0 与其预测值 $\hat{y}_0$ 之间有预测误差 $y_0-\hat{y}_0$，而 $E(y_0-\hat{y}_0)=0$，$\mathrm{var}(y_0-\hat{y}_0)=E(y_0-\hat{y}_0)^2=E(y_0-E(y_0)^2+E(E(\hat{y}_0)-\hat{y}_0)^2)$，即预测误差总方差($\sigma_F^2$)= 随机扰动产生的方差($\sigma^2$)+ 抽样误差产生的方差($\sigma_{\hat{y}_0}^2$)，扰动方差是我们控制之外的，而预测值的方差可通过增加样本容量来减少。通过计算可知

$$\sigma_F^2=\sigma^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right)$$

从而易见：如果要降低 σ_F^2，可以采取如下措施：

(1)增大样本容量 n；

(2)增大样本中自变量的分散性(即增大 l_{xx})；

(3)减少 x_0 与自变量样本均值 $\bar{x}$ 之间的距离。

前面两个结论是显然的，它们反映了总体回归直线的估计式越好，预测误差的方差越小。第三个结论使人感兴趣，它意味着对于接近 $\bar{x}$ 的自变量值来说，所作的预测要比对距离较远的自变量值更好。即我们可以凭借着某些经验和信息作出更好的预测，而经验的范围是用自变量的样本值来表示的，这个范围的中心是 $\bar{x}$。

2. 预测区间

我们用 S^2 代替 σ^2，由此可得到 σ_F^2 的一个较优良的估计

$$S_F^2=S^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right)$$

这时，我们已经给出了 $y_0-\hat{y}_0\sim N(0,\ \sigma_F^2)$，因而$\dfrac{y_0-\hat{y}_0}{\sigma_F}\sim N(0,\ 1)$，$\dfrac{y_0-\hat{y}_0}{S_F}\sim t(n-2)$。依此，我们可构造一个区间，这个区间将以某个确定的概率包含实际值 y_0。如果给定这个概率水平为 $1-\alpha$，则可以得到

$$P\left(\left|\frac{y_0-\hat{y}_0}{S_F}\right|\leqslant t_{\frac{\alpha}{2},\ n-2}\right)=1-\alpha$$

于是，y_0 的 $1-\alpha$ 预测区间为

$$(\hat{y}_0 - t_{\frac{\alpha}{2}, n-2} S_F, \ \hat{y}_0 + t_{\frac{\alpha}{2}, n-2} S_F)$$

即以 $1-\alpha$ 的概率预测 y_0 包含于上述区间内。

例 8.5.2　承例 8.5.1，求每个自变量值所对应的因变量值的预测区间。

表 8.5.3　个体预测区间

x	$\hat{y}$	95%下限	95%上限
1.2	21.11	−6.48	48.70
1.1	36.88	11.93	61.84
0.85	76.33	56.07	96.59
0.75	92.11	72.62	111.59
0.45	139.44	117.71	161.17

比如，价格为 1.1 元/斤时，黄瓜销售量的 95%的预测区间为(11.93,61.84)，即可以预测在价格 1.1 元/斤下的黄瓜销量为 12 斤至 62 斤之间。

8.6　多元回归分析

通过前几节内容介绍，我们已经明白，回归分析是对客观事物数量依存关系的分析，广泛地应用于社会经济现象变量之间的影响因素和关联的研究。同时，几乎普及应用的 Excel 软件为回归分析的求解给出了非常方便的操作过程，这样便为回归分析的实现提供了强有力的保证。

由于客观事物的联系错综复杂，经济现象的变化往往受到不只一个因素的影响。为了全面揭示这种复杂的依存关系，准确地测定现象之间的数量变动，提高预测的准确度，就要建立多元回归模型进行深入、系统的分析。

多元回归模型的一般形式为

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + \varepsilon$$

其中 y 是因变量；$x_1, x_2, \cdots, x_k$ 为 k 个自变量；ε 为随机扰动；$a, b_1, b_2, \cdots, b_k$ 为回归参数。

我们对因变量和所有自变量进行 n 次观测，得到样本数据$(y_i, x_{i1}, x_{i2}, \cdots, x_{ik})$，$i=1, 2, \cdots, n$，并假定第 i 次观测的随机误差为 ε_i，且 ε_i 服从正态分布 $N(0, \sigma^2)$，$E(\varepsilon_i \varepsilon_j)=0(i \neq j)$，于是就有

$$y_i = a + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik} + \varepsilon_i, \ i=1, 2, \cdots, n, \ \varepsilon_1, \cdots, \varepsilon_n \ iid \ N(0, \sigma^2)$$

根据最小二乘法，要求 $\min Q(a, b_1, b_2, \cdots, b_k) = \min \sum_{i=1}^{n} (y_i - E(y_i))^2$，由微积分求极值方法，需对上式的各回归系数求偏导数，并令其为零，可得到 $k+1$ 元线性方程组，其解即为回归系数 $a, b_1, b_2, \cdots, b_k$ 的最小二乘估计值 $\hat{a}, \hat{b}_1, \hat{b}_2, \cdots, \hat{b}_k$，这样就得到多元经验回归方程

$$\hat{y} = \hat{a} + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots + \hat{b}_k x_k$$

与一元回归分析类似，同样可对回归方程进行显著性检验、应用回归方程进行预测等。

上述经验回归方程的计算过程是非常复杂的，我们必须借助 Excel 软件来实现，下面我们结合实例进行介绍和分析。

例 8.6.1　某住宅小区附近的家具商城，认为住宅销售户数和新婚对数这两个因素对家具的销售额有明显的作用。为了确定该商城每季度家具的进货和销售，他们对全市各个小区家具店收集了 12 组市场调查资料如下。

表 8.6.1

家具销售额 y（万元）	住宅销售住户 x_1	结婚对数 x_2
214	114	23
248	123	22
397	239	25
388	221	24
415	248	23
430	251	26
374	232	25
550	238	27
455	254	34
535	372	39
454	247	41
638	410	38

其中，因变量 y 为季节家具销售额，x_1 为住宅销售套数，x_2 为结婚对数。请为商城人员建立二元经验回归方程并进行统计推断。

解　打开 SPSS 软件，建立如下表 8.6.2 所示数据文件（见 SPSS 数据文件例 8.6.1）。

表 8.6.2

	y	x1	x2
1	214	114	23
2	248	123	22
3	397	239	25
4	388	221	24
5	415	248	23
6	430	251	26
7	374	232	25
8	550	238	27
9	455	254	34
10	535	372	39
11	454	247	41
12	638	410	38

调用回归分析程序，依次点击：

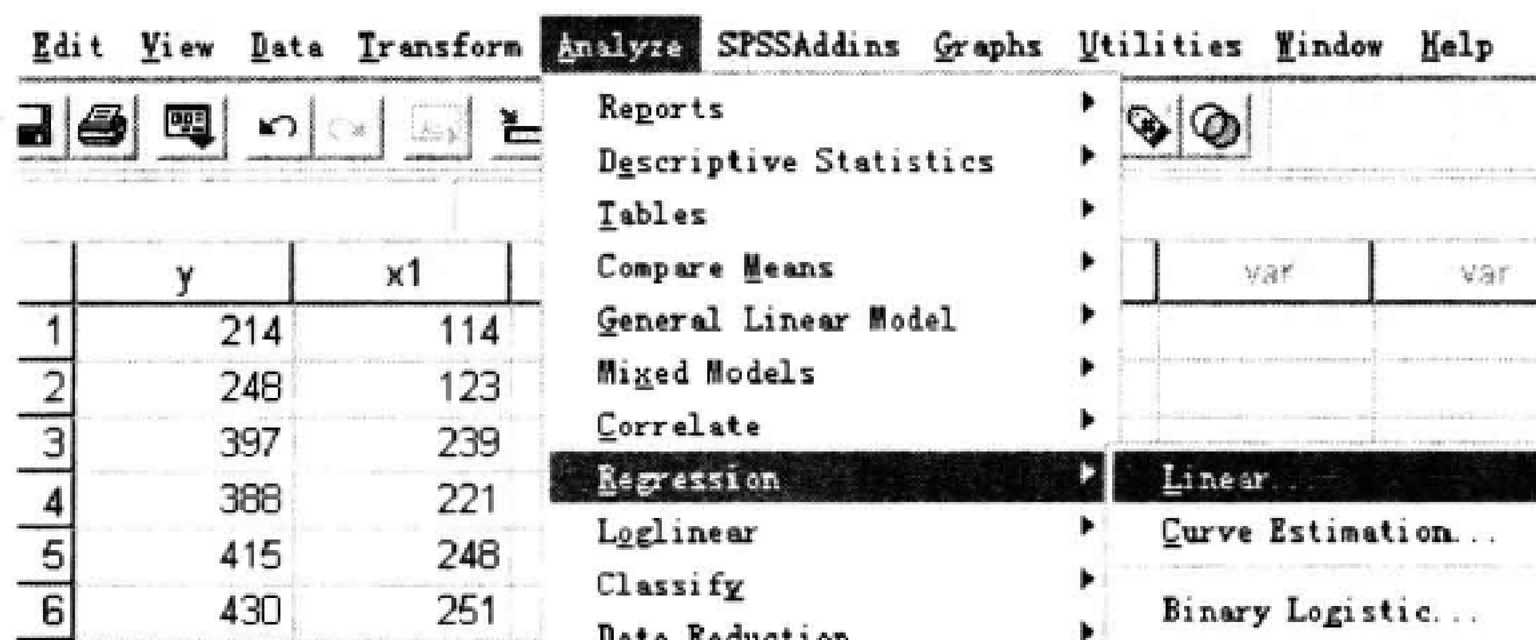


图 8.6.1

在弹出的对话框中，将因变量 y 和自变量 x_1，x_2 分别放入相应位置，直接运行即得到：

表 8.6.3 Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	X_2，X_1[a]	.	Enter

a All requested variables entered

b Dependent Variable: Y

表 8.6.4 Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.914[a]	.836	.799	53.289

a Predictors: (Constant), X_2, X_1

表 8.6.5 ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	130006.681	2	65003.340	22.891	.000[a]
	Residual	25556.986	9	2839.665		
	Total	155563.667	11			

a Predictors: (Constant), X_2, X_1

b Dependent Variable: Y

表 8.6.6 Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	89.965	68.020		1.323	.219
	X_1	1.240	.279	.869	4.451	.002
	X_2	1.039	3.311	.061	.314	.761

a Dependent Variable: Y

由此可知，所求二元经验回归方程

$$\hat{y}=89.965+1.240x_1+1.039x_2$$

并且当结婚对数保持不变时，入住户每增加一户，家具销售额平均可增加 1.240 万元，当入住户数保持不变时，每增加一对新婚数，家具销售额平均增加 1.039 万元。

对输出成果的分析，给商城人员制定销售计划有很好的指导作用。如了解到近期 20 对夫妻新婚，150 套现房出售，则本季度商城可以安排 $\hat{y}=89.965+1.240\times150+1.039\times20=296.745$ 万元的销售额。

例 8.6.2 天津某区关于“电脑销售量、人均收入和电脑平均价格”的调查资料如下。

表 8.6.7

年份 t	电脑销售量 y(台)	人均收入 x_1(元/年)	平均价格 x_2(元/台)
1994 上	1040	5789.4	13420
1994 下	1238	5989.8	12800
1995 上	1895	7096.2	12210
1995 下	2210	7396.6	11400

续表 8.6.7

年份 t	电脑销售量 y(台)	人均收入 x_1(元/年)	平均价格 x_2(元/台)
1996 上	3823	8091.2	10860
1996 下	5887	8275.3	10240
1997 上	6795	8391.6	8581
1997 下	8490	8575.7	7134
1998 上	10610	8623.5	5600

试建立电脑销售量的二元经验回归方程并进行统计推断。

解　使用 SPSS 软件，建立数据文件（见 SPSS 数据文件例 8.6.2）如下。

表 8.6.8

	y	x1	x2
1	1040	5789.4	13420
2	1238	5989.8	12800
3	1895	7096.2	12210
4	2210	7396.6	11400
5	3823	8091.2	10860
6	5887	8275.3	10240
7	6795	8391.6	8581
8	8490	8575.7	7134
9	10610	8623.5	5600

调用回归分析程序，在弹出的对话框中，将因变量 y 和自变量 x_1，x_2 分别放入相应位置，直接运行即得到：

表 8.6.9　Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	X_2，X_1[a]	.	Enter

a　All requested variables entered

b　Dependent Variable：Y

表 8.6.10　Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.986[a]	.972	.962	669.226

a　Predictors：(Constant)，X_2，X_1

表 8.6.11 ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	92610496.907	2	46305248.453	103.392	.000[a]
	Residual	2687179.093	6	447863.182		
	Total	95297676.000	8			

a Predictors: (Constant), X_2, X_1

b Dependent Variable: Y

表 8.6.12 Coefficients[a]

	Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B	
					Lower Bound	Upper Bound
	B	Std. Error				
(Constant)	16164.890	4889.622	3.306	.016	4200.417	28129.364
X_1	.147	.428	.345	.742	−.900	1.195
X_2	−1.231	.176	−7.000	.000	−1.661	−.801

a Dependent Variable: Y

由此可知,电脑销售量的二元经验回归方程为

$$\hat{y}=16164.890+0.147x_1-1.231x_2$$

判定系数 $R^2=0.972$ 表明,在电脑销售量的变化中,有 97%以上的变差可用“人均收入”和“电脑平均价格”的变化来解释,只有不到 3%的变差属于还得不到解释的随机误差。估计标准误差 $S=669.226$,S 的数值越小,表明回归方程的精度越高。在不同的样本容量条件和置信概率下,估计标准误差 S 可决定置信区间的宽度。

下面我们介绍多元线性回归方程的显著性检验。

与一元情形相同,我们是从总体中随机抽取一个样本,根据样本资料导出的多元线性经验回归方程,必须经过显著性检验,方可对总体作出结论,确证这个模型是否可靠和具有应用价值。通常的做法有 t -检验和 F -检验,t -检验是对偏回归系数 $b_1, b_2, \cdots, b_k$ 做显著性假设检验,F -检验是对多元线性回归模型做整体显著性的检验。

1. t -检验

作原假设 $H_0: b_i=0$ 备择假设 $H_1: b_i\neq 0$ $i=1, 2, \cdots, k$,检验统计量为 $t_i=$

$\frac{\hat{b}_i}{S_{\hat{b}_i}}$, $i=1, 2, \cdots, k$,若 $P(t(n-k-1)>|t_i|)<\frac{\alpha}{2}$,则拒绝原假设,说明对应的自变量 x_i 作用是显著的;反之,则接受原假设,认为该自变量的作用是不显著的。

当我们利用 Excel 求经验回归方程时,在输出成果中就已经有了这些检验结论。如对于前面的例 8.6.2 来说,从输出成果表 8.6.12 可见 $t_1=0.345$, $t_2=-7.000$,对检验水平 $\alpha=0.05$,"p - Value"即为检验统计量所对应的双尾概率,它们分别是"0.742"、"0.000423"。前者显著地大于0.025,后者显著小于 0.025,可见我们所拟合的二元线性回归方程式中的第二个自变量(平均价格)的作用是显著的,但是第一个自变量(人均收入)的影响不显著。事实上,1994 年至 1998 年,人均收入的增长远非电脑销售增长可比,同一般娱乐消费用的电器产品不同,并非收入水平的提高,而是科技的迅速发展和人们的工作、文化、教育的需求促进了电脑的销售,还有销售价格的降低也是促进因素之一。所以我们可以剔除第一个自变量,或将其换成"人均储蓄额"试作考虑。

2. F -检验

在多元回归模型中,有两个及两个以上自变量,回归模型的整体显著性是不能由任何一个偏回归系数的显著性所能替代的。因此必须采用 F -检验来判断回归模型整体的显著性。F -检验的原假设 H_0:判定系数统计量的真值等于零,检验统计量是 $F=\frac{SSR/k}{SSE/(n-k-1)}$,利用 Excel 软件,求解经验回归方程时,在输出成果中的方差分析(ANOVA),就是这一检验。当 $P(F(k, n-k-1)>F)<\alpha$ 时,就拒绝原假设,认为已建立起来的线性回归模型整体上显著有效。

如在例 8.6.2 中表 8.6.11, $F=103.392$, $P(F(2,6)>103.392)=2.242E-05\ll 0.05$,所以认为回归方程是显著有效的。

利用 SPSS 输出成果,我们还可写出回归系数的置信区间为

$$(\hat{b}_i-t_{\frac{\alpha}{2}, n-k-1}S_{\hat{b}_i}, \hat{b}_i+t_{\frac{\alpha}{2}, n-k-1}S_{\hat{b}_i})$$

如例 8.6.2 中, b_1 的置信概率 95%的区间为(−0.900,1.195), b_2 的置信概率 95%的区间为(−1.661,−0.801),由这些指标我们可以在置信概率为 95%时得到两个包络平面

$$\hat{y}_1=4200.417-0.900x_1-1.661x_2$$
$$\hat{y}_2=28129.364+1.195x_1-0.801x_2$$

可以断定,这两个包络平面所形成的空间层将包含了 95%的销售量数值。

最后,我们对多元线性回归模型的相关分析作个介绍。

多元线性回归模型中的两个或多个自变量,它们组合在一起同因变量发生一定

的依存关系，引起因变量的变化。同时，各个自变量又是相互独立的(如果经过统计检验认为不独立，则回归效果可能会变差，须另作研究)，它们同因变量的依存关系的性质和密切程度也是不同的。哪些是主要的，哪些是次要的？要解决这些问题，就需要对已建立的多元回归模型进行相关分析，即复相关分析和偏相关分析。

复相关是指一个因变量同多个自变量之间的相关关系。所有自变量共同变化时，因变量随之变化，其相关程度也可以用复相关系数来测定。复相关系数的计算指标为 R，即上述介绍的判定系数 R^2 的算术平方根。

复相关系数除表明所有自变量同因变量关系的密切程度外，同判定系数一样，也是对回归模型拟合优度的测定。如例 8.6.2 中的 $R = 0.986$，表明“人均收入”和“平均价格”作为一个整体影响因素同“电脑销售量”存在高度相关。

偏相关是指多元回归中各个自变量在其它自变量固定不变时，单个自变量同因变量的相关关系，其相关程度用偏回归系数测定(偏相关系数的计算要用 SPSS 来实现)。通过对各偏相关系数进行比较，可知某一个自变量对因变量的“净影响”。当我们所研究的客观事物本质上属于多因素影响的变量时，用多元回归、复相关和偏相关分析更为真实和精确。

练习 8

1. 研究含有必需氨基酸添加剂的某种饲料的营养价值时，用大白鼠作试验获得了关于进食量 (x) 和增重 (y)的 数据. 试分析大白鼠的进食量与增重之间有无相关。

鼠号	进食量 x (g)	增重 y(g)
1	820	165
2	780	158
3	720	130
4	867	180
5	690	134
6	787	167
7	934	186
8	679	145
9	639	120
10	820	158

2. 计算各品牌啤酒受欢迎度和口感的 Spearman 秩相关系数。

品牌编号	1	2	3	4	5	6	7	8	9	10
受欢迎程度	9	4	3	6	5	8	1	7	10	2
口感	8	2	5	4	7	9	1	6	10	3

3. 某公司下设 15 家分公司，每家分公司在一年中的销售额 x，广告费 y 和销售人员数量 z 如下表，试进行相关性分析。

	bianhao	x	y	z
1	1	7800	21	19
2	2	8400	19	20
3	3	6100	18	20
4	4	5200	15	15
5	5	9700	21	21
6	6	8900	20	19
7	7	10000	22	22
8	8	9300	24	24
9	9	6500	15	15
10	10	7300	19	18
11	11	4800	13	12
12	12	4500	11	12
13	13	6700	18	18
14	14	7500	20	19
15	15	9500	15	25

4. 在研究我国人均消费水平的问题中，把全国人均消费金额记作 y(元)，把人均国民收入记为 x(元). 我们收集到 1981—1993 年 13 年的样本数据(x_i, y_i)，$i=1, 2, \cdots, 13$。数据见下表：

年份	人均国民收入(元)	人均消费金额(元)
1981	393.8	294
1982	419.14	267
1983	460.86	289
1984	544.11	329
1985	668.29	406
1986	737.73	451
1987	859.97	513

年份	人均国民收入(元)	人均消费金额(元)
1988	1068.8	643
1989	1169.2	699
1990	1250.7	713
1991	1429.5	803
1992	1725.9	947
1993	2099.5	1148

(1) 画散点图；

(2) 用 OLS 法求出回归方程；

(3) 计算 x 与 y 的决定系数；

(4) 对回归方程作方差分析；

(5) 求出随机误差 ε 的方差 σ^2 的估计值；

(6) 给出 β_0，β_1 的 95%的区间估计；

(7) 对回归方程作残差图并做一些分析；

(8) 当国民收入增长 1 元时，用于消费大约是多少；

(9) 计算人均国民收入 2300 元时，人均消费金额是多少；

(10) 给出置信水平为 95%的近似预测区间。

5. 一家保险公司十分关心其总公司营业部加班的程度，决定认真调查一下现状。经过 10 周时间，收集了每周加班工作时间和签发的新保单数目如下表：

周序号	1	2	3	4	5	6	7	8	9	10
签发的新保单数目 x(张)	825	215	1070	550	480	920	1350	325	670	1215
加班工作时间 y(小时)	3.5	1.0	4.0	2.0	1.0	3.0	4.5	1.5	3.0	5.0

试用回归分析方法分析 y 与 x 的统计关系。

6. 一位医院管理专家声称，医院里全日制雇员的数量可以通过医院的床位数估计出来，床位数常用来度量医院的规模。一个医疗产品企业的研究人员决定建立一个回归模型，根据床位数对医院的全日制雇员数量进行预测。她调查了 12 家

医院，得到如下数据。数据按照床位数的多少排列。

床位数	全日制雇员数量	床位数	全日制雇员数量
23	69	50	138
29	95	54	178
29	10	64	156
35	118	66	184
42	126	76	176
46	125	78	225

7. 一家公司有几个分公司。公司的战略规划人员认为广告支出费用可以在一定程度上预测总销售收入。为了辅助完成长期规划，她从几个分公司搜集某年销售收入和广告费支出的数据如下（单位：百万元）。

广告费支出	销售收入
12.5	148
3.7	55
21.6	338
60.0	994
37.6	541
6.1	89
16.8	126
41.2	379

建立简单回归直线的方程，利用这些数据，根据广告费支出对销售收入进行预测。

8. 对下面的残差绘图，根据图指出回归分析的哪一个基本假定可能被违背了。

x	$y-\hat{y}$	x	$y-\hat{y}$
10	6	14	−3
11	3	15	2
12	−1	16	5
13	−11	17	8

9. 分析下面的回归分析的 SPSS 输出结果，该模型是根据 x（距消防站的距离）预测 y（火灾损失）。

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.961[a]	.923	.918	2.3163

a Predictors: (Constant)，距消防站距离

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	841.766	1	841.766	156.886	.000[a]
	Residual	69.751	13	5.365		
	Total	911.517	14			

a Predictors: (Constant)，距消防站距离

b Dependent Variable: 火灾损失

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10.278	1.420		7.237	.000
	距消防站距离	4.919	.393	.961	12.525	.000

a Dependent Variable: 火灾损失

回答下列问题：

(1)回归模型的方程是什么？

(2)x 的系数的含义是什么？

(3)回归模型的斜率的检验结果如何?

(4)对 R Square 和标准误差进行评述;

(5)对 F 值与 x 的 t 值的关系进行评述。

10. 研究下面的 SPSS 多元回归输出结果。这个模型中有几个预测变量?几个观测值?回归直线的方程是什么?根据 F 值讨论模型的效果。哪些自变量是不显著的?为什么?对模型的整体有效性进行评述。

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.993[a]	.985	.981	131.237

a Predictors: (Constant), 民航航线里程(万公里), 铁路客运量(万人), 国民收入(亿元)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.4E+07	3	4545564.831	263.922	.000[a]
	Residual	206677.3	12	17223.105		
	Total	1.4E+07	15			

a Predictors: (Constant), 民航航线里程(万公里), 铁路客运量(万人), 国民收入(亿元)

b Dependent Variable: 民航客运量(万人)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	−187.763	338.514		−.555	.589
	国民收入(亿元)	.085	.029	.589	2.895	.013
	铁路客运量(万人)	−.001	.003	−.013	−.338	.741
	民航航线里程(万公里)	16.703	8.178	.411	2.042	.064

a Dependent Variable: 民航客运量(万人)

11. 一家房地产评估公司想对某城市的房地产销售价格 y(元/平米)与地产的评估价格 x_1(万元)、房产的评估价格 x_2(万元)和使用面积 x_3(平米)建立一个模型,以便对销售价格作出合理预测。为此,收集了 20 栋住宅楼的房地产评估数据:

	bh	y	x1	x2	x3
1	1	6890	596	4497	18730
2	2	4850	900	2780	9280
3	3	5550	950	3144	11260
4	4	6200	1000	3959	12650
5	5	11650	1800	7283	22140
6	6	4500	850	2732	9120
7	7	3800	800	2986	8990
8	8	8300	2300	4775	18030
9	9	5900	810	3912	12040
10	10	4750	900	2935	17250
11	11	4050	730	4012	10800
12	12	4000	800	3168	15290
13	13	9700	2000	5851	24550
14	14	4550	800	2345	11510
15	15	4090	800	2089	11730
16	16	8000	1050	5625	19600
17	17	5600	400	2086	13440
18	18	3700	450	2261	9880
19	19	5000	340	3596	10760
20	20	2240	150	578	9620

用 SPSS 进行回归分析，回答下面的问题：

(1)这 20 栋住宅楼的平均销售价格；

(2)计算销售价格 y 与地产估价 x_1，y 与房产估价 x_2，y 与使用面积 x_3 的相关系数；

(3)在销售价格 y 的总变差中，被 x_1，x_2，x_3 所解释的比例是多少？

(4)写出 y 对 x_1，x_2，x_3 的回归方程；

(5)检验回归方程是否显著($\alpha=0.05$)？为什么？写出检验概值 p。

(6)检验各回归系数是否显著($\alpha=0.05$)？为什么？

(7)你对模型有何改进建议？对你改进的依据作出必要的说明。

(8)当地产估价为 921 万元、房产估价为 3530 万元、使用面积 13833 平米时，对销售价格、平均销售价格作出预测(点预测和区间预测)。

练习答案与提示

练习 1.1

1. $\Omega=\{$(正,正),(正,反),(反,正),(反,反)$\}$；$A=\{$(正,正),(正,反)$\}$；$B=\{$(正,正),(反,反)$\}$；$C=\{$(正,正),(正,反),(反,正)$\}$。

2. $\Omega=\{(1,1,),(1,2),\cdots,(1,6),(2,1),(2,2),\cdots,(2,6),\cdots,$
$(6,1),(6,2),\cdots,(6,6)\}$

$AB=\{(1,1),(1,3),(2,2),(3,1)\}$；

$A+B=\{(1,1),(1,3),(1,5),\cdots,(6,2),(6,4),(6,6),(1,2),$
$(2,1)\}$；

$\overline{A}C=\Phi$；$BC=\{(1,1),(2,2)\}$；

$A-B-C-D=\{(1,5),(2,4),(2,6),(4,2),(4,6),(5,1),(6,2),$
$(6,4)\}$。

3. (1)$A\overline{B}\,\overline{C}$；(2) $AB\overline{C}$；

(3)$A\overline{B}\,\overline{C}+\overline{A}B\overline{C}+\overline{A}\,\overline{B}C$；

(4) $AB\overline{C}+A\overline{B}C+\overline{A}BC$；

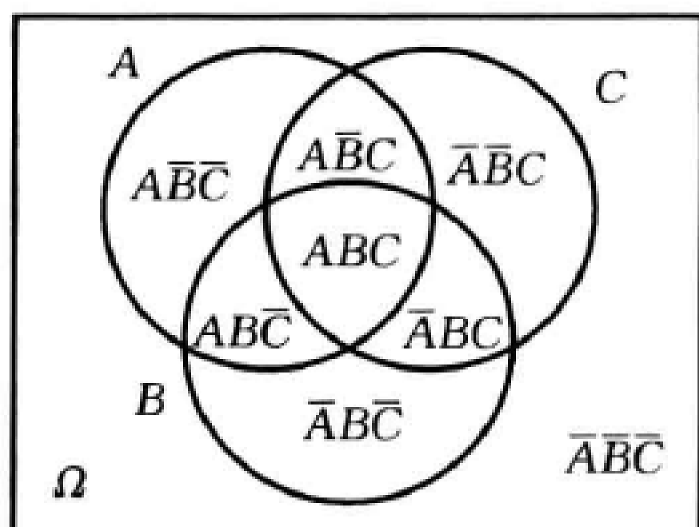
(5) $A+B+C$；(6) $\overline{A}\,\overline{B}\,\overline{C}$；

(7) $\overline{A}\,\overline{B}\,\overline{C}+\overline{A}\,\overline{B}C+A\overline{B}\,\overline{C}+\overline{A}B\overline{C}$或$\overline{A}\,\overline{B}+\overline{A}\,\overline{C}+\overline{B}\,\overline{C}$；

(8) ABC (9) $\overline{A}+\overline{B}+\overline{C}$。

4. 甲未击中；乙和丙至少一人击中；甲和乙至多有一人击中或甲和乙至少有一人未击中；甲和乙都未击中；甲和乙击中而丙未击中；甲、乙、丙三人至少有二人击中。

5. 如图



$A+B+C=A\overline{B}\,\overline{C}+A\overline{B}C+AB\overline{C}+ABC+\overline{A}BC+\overline{A}B\,\overline{C}+\overline{A}\,\overline{B}C$；

$AB+C=AB\overline{C}+C$；

$B-AC=AB\overline{C}+\overline{A}B\,\overline{C}+\overline{A}BC=B\overline{A}+AB\overline{C}=B\overline{C}+\overline{A}BC$。

6. 不一定成立。例如：$A=\{3,4,5\}$，$B=\{3\}$，$C=\{4,5\}$，那么 $A+C=B+C$，但是 $A\neq B$。

7. 不一定成立。例如：$A=\{3,4,5\}$，$B=\{4,5,6\}$，$C=\{6,7\}$，那么 $A-(B-C)=\{3\}$，但是 $(A-B)+C=\{3,6,7\}$。

练习 1.2

1. $\frac{1}{2}$；$\frac{1}{6}$；$\frac{3}{8}$。

2. $\frac{3}{8}$。

3. $P(A)=P(B)=P(C)=\frac{1}{27}$；$P(D)=P(E)=\frac{8}{27}$；$P(F)=\frac{1}{9}$；$P(G)=\frac{2}{9}$；$P(H)=\frac{8}{9}$。

4. 一次拿 3 件：(1) $\frac{C_2^1C_{98}^2}{C_{100}^3}=0.0588$；(2) $\frac{C_2^1C_{98}^2+C_{98}^1}{C_{100}^3}=0.0594$；每次拿 1 件，取后放回，拿 3 次：(1)0.0576；(2)0.0588；每次拿 1 件，取后不放回，拿 3 次：(1)0.0588；(2)0.0594.

5. $P(A_1)=\frac{C_8^3}{C_{10}^3}=\frac{7}{15}$；$P(A_2)=\frac{2C_9^3-C_8^3}{C_{10}^3}=\frac{14}{15}$ 或 $P(A_2)=1-\frac{C_8^1}{C_{10}^3}$。

6. $P=\frac{5P_9^3-4P_8^2}{P_{10}^4}=\frac{41}{90}$。

7. (1)$1-\frac{11^6}{12^6}$；(2) $\frac{C_6^4\times 11^2}{12^6}$；(3) $\frac{C_{12}^1\times C_6^4\times 11^2}{12^6}$。

8. $P=\dfrac{C_4^1C_{13}^3+C_4^1C_{13}^2C_{39}^1}{C_{52}^3}\approx 0.602$ 或 $P=\dfrac{C_{52}^3-C_4^3C_{13}^1C_{13}^1C_{13}^1}{C_{52}^3}$。

练习 1.3

1. $\dfrac{2}{3}$。 **2.** $\dfrac{1}{5}$。 **3.** (1) 0.862 (2) 0.058 (3) 0.8285。

4. (略)。 5. $P(A)=P(B)=\dfrac{1}{2}$。 **6.** (略)。 7. (略)。**8.** 0.902。

9. 系统 I：$p^n(2-p^n)$；系统 II：$p^n(2-p)^n$。

10. (1) $\dfrac{C_7^2\cdot C_3^1}{C_{10}^3}=\dfrac{21}{40}$；(2) $\dfrac{3}{10}$。

11. (1)0.10034；(2)0.0038.

12. (1) $P=C_5^2\times 0.3^2\times 0.7^2=0.3087$(贝努里试验)；(2)0.371(条件概率)。

13. (1)0.9；(2)0.887。

14. (1) $(0.94)^n$；(2) $C_n^2(0.06)^2(0.94)^{n-2}$；

(3) $1-C_n^1\times 0.06\times(0.94)^{n-1}-(0.94)^n$。

15. (1) $p(1-p)^{r-1}$；(2) $C_{r+k-1}^{r-1}p^r(1-p)^k$；(3) $C_n^r p^r(1-p)^{n-r}$；

(4) $C_{n-1}^{r-1}p^r(1-p)^{n-r}$。

16. 0.458。

练习 2.2

1. (1)是概率分布；(2) $\dfrac{1}{3}$，$\dfrac{1}{16}$。

2. $C=(1-e^{-\lambda})^{-1}$。

3. $P(X=k)=p(1-p)^{k-1}$，$k=1,2,\cdots$

4. (1) $P(X=k)=(1-p)^k p$，$k=0,1,2,\cdots$；

(2) $P(X\geqslant 5)=\sum\limits_{k=5}^{\infty}P(X=k)=0.9^5=0.590$。

5. $\dfrac{1}{64}$。

6. (1) $1-0.99^{20}-20\times 0.01\times 0.99^{19}\approx 0.0175$(按 Poisson 分布近似)；

(2)$n=100$，$np=100\times 0.01=1=\lambda$(按 Poisson 分布近似)，

$$P(X \geqslant N+1) = \sum_{k=N+1}^{100} C_{100}^{k} \times 0.01^{k} \times 0.99^{100-k}$$
$$\approx \sum_{k=N+1}^{\infty} \frac{1^{k} \times e^{-1}}{k!} \leqslant 0.01$$

查表得 $N = 4$。

7. (1) $\lambda = \ln 2$；(2) $P(X > 1) = 1 - \frac{1}{2} - \frac{1}{2}\ln 2$。

8. e^{-8}。

9. (1) $e^{-\frac{3}{2}}$；(2) $1 - e^{-\frac{5}{2}}$。

10. (1) $a = \frac{1}{10}$；(2) 如下表

X	-1	0	3	8
P	$\frac{3}{10}$	$\frac{1}{5}$	$\frac{3}{10}$	$\frac{1}{5}$

练习 2.3

1. (1) $t = -1$；(2) $f(x) = \begin{cases} \frac{1}{2}x + \frac{1}{2}, & x \in [-1, 0) \\ -\frac{1}{6}x + \frac{1}{2}, & x \in [0, 3) \\ 0, & \text{其它} \end{cases}$

(3) $P(-2 < X \leqslant 2) = \frac{11}{12}$。

2. (1) $a = \frac{\pi}{2}$；$P(X > \frac{\pi}{6}) = \frac{\sqrt{3}}{2}$。

3. $\frac{1}{\sqrt{\pi}} e^{-\frac{1}{4}}$。

4. $\mu = 2$，$\sigma^2 = 3$，$C = 2$。

5. $P(Y = 2) = \frac{9}{64}$。

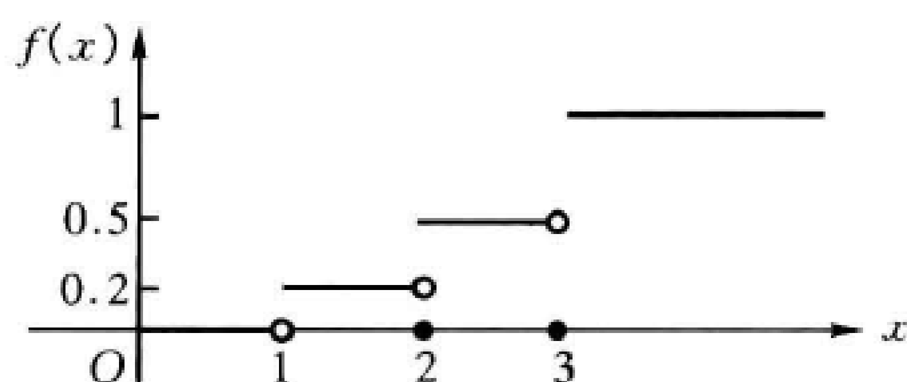
6. (1) $\frac{1}{4}(x_2 - 1)$；(2) $\frac{1}{4}(5 - x_1)$。

7. (1) $P(Y = k) = C_5^k e^{-2k}(1 - e^{-2})^{5-k}$，$k = 0, 1, 2, 3, 4, 5$；

(2) $P(Y \geqslant 1) = 1 - (1 - e^{-2})^5 \approx 0.5167$。

练习 2.4

1. $F(x)=\begin{cases}0, & x<1\\ 0.2, & 1\leqslant x<2;\\ 0.5, & 2\leqslant x<3;\\ 1, & x\geqslant 3\end{cases}$ $P(0.5\leqslant X\leqslant 2)=0.5$



2. (1) 如下表；(2) $\dfrac{2}{3}$。

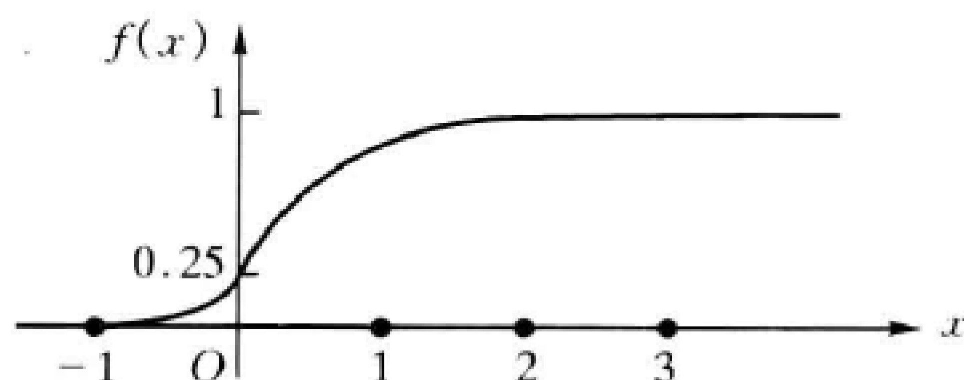
X	-1	1	3
P	0.4	0.4	0.2

3. (1) 如下表：

X	0	1	2	3
P	$\frac{27}{125}$	$\frac{54}{125}$	$\frac{36}{125}$	$\frac{8}{125}$

(2) $F(x)=\begin{cases}0, & x<0\\ \dfrac{27}{125}, & 0\leqslant x<1\\ \dfrac{81}{125}, & 1\leqslant x<2\\ \dfrac{117}{125}, & 2\leqslant x<3\\ 1, & x\geqslant 3\end{cases}$

4. $F(x)=\begin{cases}0, & x<-1\\ \frac{1}{4}x^2+\frac{1}{2}x+\frac{1}{4}, & -1\leqslant x<0\\ -\frac{1}{12}x^2+\frac{1}{2}x+\frac{1}{4}, & 0\leqslant x<3\\ 1, & x\geqslant 3\end{cases}$



5. (1) $A=1$，$B=-1$；(2) $1-e^{-2}$；(3) $f(x)=\begin{cases}2e^{-2x}, & x>0\\ 0, & x\leqslant 0\end{cases}$

6. $a=0$，$b=1$，$c=-1$，$d=1$。

7. $a=1$；$F(x)=\frac{1}{2}+\frac{1}{\pi}\arctan x$；$P(|X|<1)=0.5$。

8. (1) $P(X>t)=P(N(t)=0)=e^{-0.1t}$ 得 $F(x)=\begin{cases}1-e^{-0.1x}, & x>0\\ 0, & x\leqslant 0\end{cases}$

(2) $F(3)=0.26$；(3) $F(5)-F(3)=0.13$。

9. (1)0.8051；(2)0.5498；(3)0.6678；(4)0.8253。

10. 79.6 分。 **11.**（略）。

12. $f_Y(y)=\begin{cases}f_X\left[\frac{y-d}{c}\right]\cdot\frac{1}{|c|}, & a\leqslant\frac{|y-d|}{c}\leqslant b\\ 0, & \text{其它}\end{cases}$

当 $c>0$ 时，有 $f_Y(y)=\begin{cases}\frac{1}{c(b-a)}, & ca+d\leqslant y\leqslant cb+d\\ 0, & \text{其它}\end{cases}$

当 $c<0$ 时，有 $f_Y(y)=\begin{cases}-\frac{1}{c(b-a)}, & cb+d\leqslant y\leqslant ca+d;\\ 0, & \text{其它}\end{cases}$

练习 3.1

1.

X \ Y	0	1	$p_{i\cdot}$
0	$\frac{4}{25}$	$\frac{6}{25}$	$\frac{10}{25}$
1	$\frac{6}{25}$	$\frac{9}{25}$	$\frac{15}{25}$
$p_{\cdot j}$	$\frac{10}{25}$	$\frac{15}{25}$	1

有放回的情形

X \ Y	0	1	$p_{i\cdot}$
0	$\frac{2}{15}$	$\frac{4}{15}$	$\frac{6}{15}$
1	$\frac{4}{15}$	$\frac{5}{15}$	$\frac{9}{15}$
$p_{\cdot j}$	$\frac{6}{25}$	$\frac{9}{25}$	1

无放回的情形

2.

X \ Y	0	1	2	$p_{i\cdot}$
0	$\frac{1}{120}$	$\frac{1}{20}$	$\frac{1}{40}$	$\frac{1}{12}$
1	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{24}$	$\frac{5}{12}$
2	$\frac{1}{4}$	$\frac{1}{6}$	0	$\frac{5}{12}$
3	$\frac{1}{12}$	0	0	$\frac{1}{12}$
$p_{\cdot j}$	$\frac{7}{15}$	$\frac{7}{15}$	$\frac{1}{15}$	1

3. (1)

X \ Y	−1	0	1	$p_{i\cdot}$
0	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{1}{2}$
1	0	$\frac{1}{2}$	0	$\frac{1}{2}$
$p_{\cdot j}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

(2) $P(X=Y)=0$。

4. (1) $A=2$； (2) $f_1(x)=\begin{cases} e^{-x}, & x>0 \\ 0, & x\leqslant 0 \end{cases}$，$f_2(y)=\begin{cases} 2e^{-2y}, & y>0 \\ 0, & y\leqslant 0 \end{cases}$

(3) $(1-e^{-2})(1-e^{-6})$；(4) $1-2e^{-1}$；(5) $\frac{1}{3}$。

5. (1) $f_1(x)=\begin{cases} 2x^2+\frac{2}{3}x, & 0\leqslant x\leqslant 1 \\ 0, & \text{其它} \end{cases}$

$$f_2(y)=\begin{cases} \frac{1}{3}+\frac{1}{6}y, & 0\leqslant y\leqslant 2 \\ 0, & \text{其它} \end{cases}$$

(2) $\frac{5}{32}$。

6. (1) $f_1(x)=\begin{cases} e^{-x}, & x>0 \\ 0, & x\leqslant 0 \end{cases}$ $f_2(y)=\begin{cases} ye^{-y}, & y>0 \\ 0, & y\leqslant 0 \end{cases}$

(2) $e^{-2}-3e^{-4}$。

7. 提示：利用二维均匀分布计算得$\frac{5}{9}$。

练习 3.2

1. A。

2.

X	-1	0	1
P	0.1344	0.7312	0.1344

3. U 与 V 的联合概率分布见下表

X \ Y	0	1
0	$\frac{1}{4}$	0
1	$\frac{1}{4}$	$\frac{1}{2}$

4. $F(x, y)=\begin{cases}(1-e^{-x})(1-e^{-2y}), & x>0, y>0 \\ 0, & \text{其它}\end{cases}$

$$F(x, y)=\begin{cases}0 & x<0 \text{ 或 } y<0 \\ \frac{1}{3}x^2 y(x+\frac{y}{4}) & 0\leqslant x<1, 0\leqslant y<2 \\ \frac{1}{3}x^2(2x+1) & 0\leqslant x<1, y\geqslant 2 \\ \frac{1}{12}y(4+y) & x\geqslant 1, 0\leqslant y<2 \\ 1 & x\geqslant 1, y\geqslant 2\end{cases}$$

$$F(x, y)=\begin{cases}0, & x<0 \text{ 或 } y<0 \\ 1-e^{-x}-xe^{-y}, & 0\leqslant x<y \\ 1-e^{-y}-ye^{-y}, & 0\leqslant y\leqslant x\end{cases}$$

5. $F(x, y)=\begin{cases}0, & x<0 \text{ 或 } y<0 \\ x^2y^2, & 0\leqslant x<1, 0\leqslant y<1 \\ x^2, & 0\leqslant x<1, y\geqslant 1 \\ y^2, & x\geqslant 1, 0\leqslant y<1 \\ 1, & x\geqslant 1, y\geqslant 1\end{cases}$

6. (1) $A=1$ (2) $f_Z(z)=\begin{cases}0, & z<0 \\ \frac{1}{2}(1-e^{-z}), & 0\leqslant z<2 \\ \frac{1}{2}(e^2-1)e^{-z}, & z\geqslant 2\end{cases}$

7. (1) $A=\frac{1}{\pi^2}$, $B=C=\frac{\pi}{2}$；(2) $f(x, y)=\frac{12}{\pi^2(x^2+9)(y^2+16)}$;

(3) $F_X(x)=\frac{1}{2}+\frac{1}{\pi}\arctan\frac{x}{3}$, $F_Y(y)=\frac{1}{2}+\frac{1}{\pi}\arctan\frac{y}{4}$;

$f_X(x)=\frac{3}{\pi^2(x^2+9)}$, $f_Y(y)=\frac{4}{\pi^2(x^2+16)}$。

(4) $\frac{3}{4}$, $\frac{3}{4}$, $\frac{9}{16}$；(5) X 与 Y 相互独立。

8. (1) $F_X(x)=\begin{cases}1-e^{-0.5x}, & x>0 \\ 0, & x\leqslant 0\end{cases}$ $F_Y(y)=\begin{cases}1-e^{-0.5y}, & y>0 \\ 0, & y\leqslant 0\end{cases}$

(2) $f(x, y)=\begin{cases}0.25e^{-0.5(x+y)}, & x>0, y>0 \\ 0, & \text{其它}\end{cases}$

$$f_X(x)=\begin{cases}0.5e^{-0.5x}, & x>0\\ 0, & x\leqslant 0\end{cases}\quad f_Y(y)=\begin{cases}0.5e^{-0.5y}, & y>0\\ 0, & y\leqslant 0\end{cases}$$

(3) X 与 Y 相互独立；

(4) $e^{-0.1}(\approx 0.9048)$。

9. (1) $f(x,y)=\begin{cases}12e^{-3x-4y}, & x>0,\ y>0\\ 0, & 其它\end{cases}$

$$F(x,y)=\begin{cases}(1-e^{-3x})(1-e^{-4y}), & x>0,\ y>0\\ 0, & 其它\end{cases}$$

(2) $(1-e^{-3})(1-e^{-4})$；(3) $1-4e^{-3}$。

10. $\frac{5}{7}$，$\frac{4}{7}$。

11. $f_z(z)=\begin{cases}\frac{3}{2}(1-z^2), & 0\leqslant z<1\\ 0, & 其它\end{cases}$

练习 3.3

1.

$Y \mid X=2$	1	2	3
P	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

2. 当 $x>0$ 时，$f(y\mid x)=\begin{cases}\frac{1}{x}, & 0<y<x\\ 0, & y\leqslant 0 或 y\geqslant x\end{cases}$

3. 当 $-1<y<1$ 时，$f(x\mid y)=\begin{cases}\frac{1}{1-\mid y\mid}, & \mid y\mid<x<1\\ 0, & 其它\end{cases}$

练习 4.1

1. $E(X)=0.5$，$E(X)=0.5$。

2. $D(X)=\frac{3}{8}$，$D(X)=\frac{27}{76}$。

3. $x_3=21$，$a=0.2$。

4. $a < b < \dfrac{a}{1-p}$，对于 m 个人可期望获益 $ma - mb(1-p)$。

5. $k=3$，$\alpha=2$。

6. $E(X)=1$，$D(X)=\dfrac{1}{6}$。

7. $E(Z)=\dfrac{2+\sigma}{\sigma}$，$D(Z)=\dfrac{2+9\sigma^4}{\sigma^2}$。

8. (1) $E(X_1)=2$，$D(X_1)=4$　(2) $c=\dfrac{1}{4}$，$E(X_2)=4$。

9. $D(XY)=27$。

10. $E(X)=270$，$D(X)=225$。

11. $0, 1, \cdots, 9$；$P(X\leqslant 8)=0.9999$。

12. $\dfrac{\pi}{24}(a+b)(a^2+b^2)$。

13. $a=12$，$b=-12$，$c=3$。

14. $E(XY)=4$。

15. 提示：$f(x)=\begin{cases}\dfrac{1}{3}, & -1\leqslant x\leqslant 2\\ 0, & \text{其它}\end{cases}$，于是 $P(Y=1)=P(X>0)=\dfrac{2}{3}$，$P(Y=-1)=P(X<0)=\dfrac{1}{3}$，$P(Y=0)=0$。$D(Y)=\dfrac{8}{9}$。

16. 提示：设每周的产量为 N，显然 $N\leqslant 5$，每周利润

$$L=\begin{cases}(c_2-c_1)N, & Q>N\\ c_2Q-c_1N-c_3(N-Q), & Q\leqslant N\end{cases}=\begin{cases}6N, & Q>N\\ 10Q-4N, & Q\leqslant N\end{cases}$$

$$\begin{aligned}E(L)&=6N\cdot P(Q>N)+(10Q-4N)\cdot P(Q\leqslant N)\\&=6N\cdot\sum_{n=N}^{5}\frac{1}{5}+10\sum_{n=1}^{N}n\cdot\frac{1}{5}-4N\cdot\sum_{n=1}^{N}\frac{1}{5}\\&=\frac{6}{5}N(5-N)+2\cdot\frac{1+N}{2}\cdot N-\frac{4N^2}{5}\\&=7N-N^2\end{aligned}$$

令 $\dfrac{\mathrm{d}}{\mathrm{d}N}(E(L))=7-2N=0$ 得 $N=3.5$。又因 $\dfrac{\mathrm{d}^2}{\mathrm{d}N^2}(E(L))=-2<0$，所以当 $N=3.5$ 时，$E(N)$ 达到最大值。

由于 Q 和 N 均取正整数，所以应取 $N=3$ 或 $N=4$，故当产量为 3 件或 4 件时，利润达到最大期望值 12 元。

17. $P(400<X<600)\geqslant\dfrac{39}{40}$。

18. 提示:先猜 A 类题的期望为

$$0\times(1-p)+ap(1-p)+(a+b)pq;$$

先猜 B 类题的期望为

$$0\times(1-q)+aq(1-q)+(a+b)qp$$

所以若 $ap(1-p)\geqslant bq(1-q)$,则应先猜 A 类题,否则先猜 B 类题。

19. 2。

20. $9\left[1-\left(\frac{8}{9}\right)^{25}\right]$。

练习 4.2

1. (1) $(0, 0)$, $(\frac{3}{4}, \frac{3}{4})$;(2) $\mathrm{cov}(X, Y)=0$, $\rho_{XY}=0$; (3) X 与 Y 不相关,但也不独立。

2. (1) $C=8$; (2) $(\frac{4}{5}, \frac{8}{15})$, $(\frac{2}{75}, \frac{11}{225})$;

(3) $\mathrm{cov}(X, Y)=\frac{4}{225}$, $\rho_{XY}=\frac{2\sqrt{66}}{33}$;(4) X 与 Y 相关,不独立。

3. $D(X+Y)=85$, $D(X-Y)=37$。

4. $\rho=\pm\rho_{XY}$。

5. $E(Y)=4$, $D(Y)=18$, $\mathrm{cov}(X, Y)=6$, $\rho_{XY}=1$。

6. $E(Y)=\frac{4}{3}$, $D(Y)=\frac{29}{45}$, $\mathrm{cov}(X, Y)=\frac{7}{9}$, $\rho_{XY}=\frac{7}{3}\sqrt{\frac{5}{29}}$。

7. $\rho_{Z_1Z_2}=\frac{\alpha^2-\beta^2}{\alpha^2+\beta^2}$。

8. $a=\pm1$, $b=\frac{1}{\sqrt{3}}$, $c=-\frac{2}{\sqrt{3}}$; $a=\pm1$, $b=-\frac{1}{\sqrt{3}}$, $c=\frac{2}{\sqrt{3}}$。

9. $P\{|X+Y|\geqslant 6\}\leqslant\frac{1}{12}$。

练习 4.3

1. $P(X>102)\approx0.0228$

2. $P(180\leqslant X\leqslant 220)\approx0.8764$

3. 提示:假设第 i 个加数的取整误差为 $X_i(i=1, 2, \cdots, n)$,则 $E(X_i)=0$,

$D(X_i)=\frac{1}{12}$。令 $S_n=\sum_{i=1}^{n}X_i$，当 n 足够大时，$S_n \sim N(0, \frac{n}{12})$。(1) 0.0026；(2) $n \leqslant$ 446.16，$n=446$；(3) $[-14.85, 14.85]$。

4. $P(T>360)=0.1367$。

5. $P(X>10)=0.045$

6. (1) $n \geqslant 5250$；(2) $n \geqslant 1423.5$

7. 提示：X 为一年内死亡的人数，$X \sim B(10000, 0.006)$，公司利润为 $Y=10000\times 12-1000X$。(1) 约为 0；(2) 约为 0.5。

8. 2265 单位

9. $n=147$。

练习 5.1

1. 0，0.01，0.5，0.8414

2. 106　**3.** 5.43　**4.** 0.95　**5.** 26.105

6. 0.6826，0.8426，0.90

7. 0.1　**8.** B　9. $F(1, n)$　**10.** $\frac{1}{8}$，$\frac{1}{12}$，$\frac{1}{16}$，3

11. $\chi^2(n-k)$

12. $t(9)$　**15.** $F(10, 5)$　**16.** $F(5, n-5)$

练习 5.2

3. $\hat{\mu}_2=\frac{1}{3}X_1+\frac{1}{3}X_2+\frac{1}{3}X_3$ 最有效

4. 2.68　**5.** 1143.75，96.06

6. $a=\frac{n_1}{n_1+n_2}$，$b=\frac{n_2}{n_1+n_2}$

练习 5.3

1. (4.412，5.588)

2. ①457.50　② (432.30，482.70)

③ (438.90，476.09)　④1240.28

⑤ (586.79 ，4134.26)　⑥ 35.2176

⑦ (24.2237 , 64.2982)

3. $\left[\frac{15.37\sigma^2}{k^2}\right]+1$　　**4.** ① $e^{\mu+\frac{1}{2}}$　　② $(-0.98, 0.98)$　　③ $(e^{-0.48}, e^{1.48})$

5. $(3.42, 8.58)$

6. ① $(0.03822, 1.635265)$　　② $(-0.387, 0.427)$

练习 5.4

1. θ 的矩估计量为 $\hat{\theta}_1=\frac{2}{\pi}\overline{X}^2$，$\hat{\theta}_1$ 不是 θ 的无偏估计量；θ 的最大似然估计量为 $\hat{\theta}_2=\frac{1}{2n}\sum\limits_{i=1}^{n}X_i^2$，$\hat{\theta}_2$ 是 θ 的无偏估计量。

2. $\hat{\mu}=\frac{1}{n}\sum\limits_{i=1}^{n}\ln X_i$，$\hat{\sigma}^2=\frac{1}{n}\sum\limits_{i=1}^{n}(\ln X_i-\hat{\mu})^2$，因为 $E(X)=e^{\mu+\frac{1}{2}\sigma^2}$，所以 $E(X)=e^{\mu+\frac{1}{2}\sigma^2}$ 的最大似然估计为 $e^{\hat{\mu}+\frac{1}{2}\hat{\sigma}^2}$。

3. (1) 因为 $E(X)=D(X)=\theta$，所以 θ 的矩估计量有三个：$\hat{\theta}_1=\overline{X}$，$\hat{\theta}_2=B_2$，$\hat{\theta}_3=S^2$；

(2) θ 的最大似然估计量为 $\hat{\theta}_1=\overline{X}$；

(3) $\hat{\theta}_1=\overline{X}$ 与 $\hat{\theta}_3=S^2$ 均为 θ 的无偏估计量.

练习 6.2

1. $\frac{\overline{X}}{Q}\sqrt{n(n-1)}$。　**2.** 能认为这批钢索的断裂强度为800（千克力 / 平方厘米）

3. 可以认为机器工作是正常的。　　4. 能。

5. 没有显著变化。　　6. 工作不正常。

练习 6.3

1. 首先检验两正态总体的方差是否相等，再检验乙总体的均值是否比甲总体的均值大. 通过双侧检验，可认为两总体的方差是相等的；通过单侧检验，认为采用乙方案可以比甲方案提高得率。

2. (1) 方差相等；(2) 均值不相等。

3. 注意训练前后的分数是不独立的，应该使用成对数据比较检验法，通过检验，认为体能训练效果不显著。